

# Gemini: Graph estimation with matrix variate normal instances

Shuheng Zhou,  
Department of Statistics,  
University of Michigan, Ann Arbor, MI 48109

Technical Report #531

September 23, 2012

## Abstract

Undirected graphs can be used to describe matrix variate distributions. In this paper, we develop new methods for estimating the graphical structures and underlying parameters, namely, the row and column covariance and inverse covariance matrices from the matrix variate data. Under sparsity conditions, we show that one is able to recover the graphs and covariance matrices with a single random matrix from the matrix variate normal distribution. We establish consistency and obtain the rates of convergence in the operator and the Frobenius norm. We then extend this analysis to the multiple-sample instances, and show that having replicates will allow one to estimate more complicated graphical structures and achieve faster rates of convergence in both norms. We provide simulation evidence showing that we can recover graphical structures as well as estimating the precision matrices, as predicted by theory.

**Keywords.** Graphical model selection, covariance estimation, inverse covariance estimation, Graphical Lasso, Matrix variate normal distribution.

## 1 Introduction

The matrix variate normal model has a long history in psychology and social sciences, and is becoming increasingly popular in biology and genetics, econometric theory, image and signal processing, and machine learning in recent years. In this paper we present a theoretical framework to show that one can estimate the covariance and inverse covariance matrices well using only one matrix from the matrix-variate normal distribution. The motivation for this problem comes from many applications in statistics and machine learning. For example in microarray studies, a single  $f \times m$  data matrix  $X$  represents expression levels for  $m$  genes on  $f$  microarrays; one needs to find out simultaneously the correlations and partial correlations between genes, as well as between microarrays. Another example concerns observations from a spatio-temporal stochastic process which can be described with a matrix normal distribution with a separable covariance matrix  $S \otimes T$ , where typically,  $S$  is called spatial covariance,  $T$  is called the temporal covariance, and  $\otimes$  is the Kronecker product. When the stochastic process is spatio-temporal, some structures can be assumed for one or both of the matrices in the Kronecker product. However, typically one has only one observational matrix.

We call the random matrix  $X$  which contains  $f$  rows and  $m$  columns a single data matrix, or one instance from the matrix variate normal distribution. We say that an  $f \times m$  random matrix  $X$  follows a matrix normal

distribution with a separable covariance matrix  $\Sigma = A \otimes B$ , which we write

$$X_{f \times m} \sim \mathcal{N}_{f,m}(M, A_{m \times m} \otimes B_{f \times f}), \quad (1)$$

is equivalent to say  $\text{vec}\{X\}$  follows a multivariate normal distribution with mean  $\text{vec}\{M\}$  and covariance  $\Sigma = A \otimes B$ . Here  $\text{vec}\{X\}$  is formed by stacking the columns of  $X$  into a vector in  $\mathbf{R}^{mf}$ . Intuitively,  $A$  describes the covariance between columns of  $X$  while  $B$  describes the covariance between rows of  $X$ . See [Gupta and Varga \(1992\)](#); [Dawid \(1981\)](#) for characterization and examples. Note that we can only estimate  $A$  and  $B$  up to a scaled factor, as  $A\eta \otimes \frac{1}{\eta}B = A \otimes B$  for any  $\eta > 0$ , and hence this will be our goal of the paper, and precisely what we mean, when we say we are interested in estimating covariances  $A$  and  $B$ .

Undirected graphical models are often used to describe high dimensional distributions. We will use such descriptions in the present work to encode structural assumptions on the inverse of the row and column covariance matrices. A common structural assumption is that the inverse covariance matrices, also known as the precision matrices, are sparse, which means that the number of nonzero entries (sparsity levels) in one or both of them are bounded. Under sparsity assumptions, a popular approach to obtain a sparse estimate for the precision matrix is given by the  $\ell_1$ -norm regularized maximum-likelihood function, also known as the GLasso [Yuan and Lin \(2007\)](#); [Friedman et al. \(2008\)](#); [Banerjee et al. \(2008\)](#); [Rothman et al. \(2008\)](#). All these methods and their analysis assume that one is given independent samples and the estimation of  $A$  or  $B$  alone was their primary goal, as they all assume that  $X$  has either independent rows or independent columns. A direct application of the GLasso estimator to estimate  $A \otimes B$  with no regard for its separable structure will lead to computational misery, as the cost will become prohibitive for  $f, m$  in the order of 100 (c.f. Section 2.4). A mean-restricted matrix-variate normal model was considered in [Allen and Tibshirani \(2010\)](#), where they proposed placing additive penalties on estimated inverse covariance matrices in order to obtain regularized row and column covariance/precision matrices. The focus of their work is on missing value imputation, for example, when the data is given by the Netflix movie rating data, rather than estimation of the graphs or the underlying parameters. We review this and other closely related work in Section 3.2.

In this work, we take a penalized approach and show from a theoretical point of view, the advantages of treating the estimation of covariance matrices  $A$ ,  $B$ , and the graphs corresponding to their inverses simultaneously albeit through separable optimization functions. The key observation and starting point of our work is: although  $A$  and  $B$  are not identifiable given the separable representation as in (1), their correlation matrices  $\rho(A)$  and  $\rho(B)$ , and the graphical structures corresponding to their inverses are identifiable, and can indeed be efficiently estimated for a given  $X \sim \mathcal{N}_{f,m}(0, A \otimes B)$ . Moreover, in the mean-zero matrix normal model,  $\rho(A)^{-1}$  and  $\rho(B)^{-1}$  encode the same amount of structural information as  $A^{-1}$  and  $B^{-1}$  do, in the sense that they share an identical set of non-zero edges. Therefore, we propose estimating the overall  $\Sigma = A \otimes B$  and its inverse by (1) first estimating correlation matrices  $\rho(A)$  and  $\rho(B)$  (and their inverses) simultaneously using a pair of  $\ell_1$ -norm penalized estimators for an instance  $X \sim \mathcal{N}_{f,m}(0, A \otimes B)$ , (2) and then combining these two estimators with the estimated variances to form an estimator for  $\Sigma$ .

**Contributions.** In this paper, we will answer the following question: how sparse does  $A^{-1}$  or  $B^{-1}$  need to be in order for us to obtain statistical convergence rates in terms of the operator norm and the Frobenius norm for estimating  $A$  and  $B$  (up to a scaled factor) simultaneously with one data matrix  $X$ ? Toward this end, we develop Gemini, a new method for estimating graphical structures, and the underlying parameters  $A$  and  $B$ , in a mean-zero matrix variate normal model. Under suitable assumptions, we establish convergence bounds in the operator and the Frobenius norm for estimating both  $A, B$  and their inverses. Our estimators and convergence analysis extend, with suitable adaptation, to the general setting where  $n$  replicates of  $X$

are available, for which with all other parameters hold invariant, the rates of convergence for estimating  $A$ ,  $B$ , and their inverses will be proportional to  $n^{-1/2}$ . In addition, we provide simulation evidence that we can recover graphical structures as well as estimate the precision matrices effectively.

In summary, we make the following theoretical contributions: (i) consistency and rates of convergence in the operator and the Frobenius norm of the covariance matrices and their inverses, (ii) large deviation results for the sample correlation estimators which we propose for estimating both the row and column correlations given a single matrix or multiple replicates of the matrix-normal data, (iii) conditions that guarantee simultaneous estimation of the graphs for both rows and columns. To the best of our knowledge, these are the first such results on the matrix-variate normal distributions in the high dimensional setting for finite sample instances, by which we mean  $n < \log \max(m, f)$ . Also worthy of mentioning is the computational efficiency of our method. The dominating costs involve in estimating  $\rho(A)^{-1}$  and  $\rho(B)^{-1}$ , the total cost of which is in the order of  $O(f^3 + m^3)$  for sparse graphs or  $O(f^4 + m^4)$  for general graphs.

There is no known closed-form solution for the maximum of the likelihood function for the matrix-variate normal distribution (cf. (9)). There has been a line of work in the literature which suggested using iterative algorithms, namely, the flip-flop methods to estimate the covariance matrix with the Kronecker structure; see for example Dutilleul (1999); Lu and Zimmerman (2005); Werner et al. (2008) and references therein. In the present work, building upon the baseline Gemini estimators, we also propose a three-step penalized variant of the flip-flop algorithms in Section 5. We show that under certain conditions, this approach yields improvements upon the baseline Gemini estimators.

The rest of the paper is organized as follows. In Section 2, we will define our model and the method. Section 3 presents the main theoretical results in this paper on estimating  $A \otimes B$ , as well as discussions on our method and results; moreover, we review the related work on covariance estimation and Gaussian graphical model selection, to place our work in context. Section 4 provides large deviation inequalities for the sample correlation coefficients in approximating the underlying parameters of  $\rho(A)$  and  $\rho(B)$ . Moreover, convergence rates in the Frobenius norm for estimating the inverse correlation matrices are derived. We propose a non-iterative penalized flip-flop algorithm and study its convergence properties in Section 5 and 6. Section 7 shows our numerical results. We conclude in Section 8. We place all technical proofs in the Appendix.

## 1.1 Notation

For a matrix  $A$ , let  $|A|_{\max} = \max_{i,j} |A_{ij}|$  denote the element-wise or entry-wise max norm. Let  $|A|$  denote the determinant and  $\text{tr}(A)$  be the trace of  $A$ . Let  $\varphi_{\max}(A)$  and  $\varphi_{\min}(A)$  be the largest and smallest eigenvalues, respectively. We use  $A^{-T}$  as a shorthand notation for  $(A^{-1})^T$ . We write  $|\cdot|_1$  for the  $\ell_1$  norm of a matrix vectorized, i.e., for a matrix  $|A|_1 = \|\text{vec}\{A\}\|_1 = \sum_i \sum_j |A_{ij}|$ . Let  $\kappa(A)$  denote the condition number for matrix  $A$ . We write  $|A|_{1,\text{off}} = \sum_{i \neq j} |A_{ij}|$ , and write  $|A|_{0,\text{off}}$  for the number of non-zero non-diagonal entries in the matrix. We let  $C$  be a constant which may change from line to line. For two numbers  $a, b$ ,  $a \wedge b := \min(a, b)$ , and  $a \vee b := \max(a, b)$ . We write  $\text{diag}(A)$  for a diagonal matrix with the same diagonal as  $A$ . The matrix Frobenius norm is given by  $\|W\|_F = \sqrt{\sum_i \sum_j w_{ij}^2}$ . The operator norm  $\|W\|_2^2$  is given by  $\varphi_{\max}(WW^T)$ . For a symmetric matrix  $A$ , let  $\Upsilon(A) = (v_{ij})$  where  $v_{ij} = \mathbb{I}_{a_{ij} \neq 0}$ , where  $\mathbb{I}_{\{\cdot\}}$  is the indicator function. We write  $a \asymp b$  if  $ca \leq b \leq Ca$  for some positive absolute constants  $c, C$  which are independent of  $n, f, m$  or sparsity parameters.

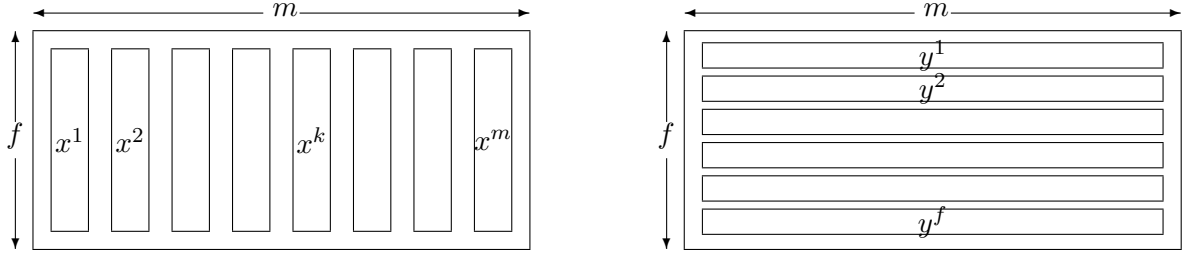


Figure 1: Column and row vectors of matrix  $X$ . The normalized column vectors  $x^1/\sqrt{a_{11}}, \dots, x^m/\sqrt{a_{mm}}$ , where  $a_{ii} > 0, \forall i$  follow a multivariate normal distribution  $N_f(0, B)$  while normalized column vectors  $y^1/\sqrt{b_{11}}, \dots, y^f/\sqrt{b_{ff}}$ , where  $b_{jj} > 0$ , for  $Y = X^T$  follow a multivariate normal distribution  $N_m(0, A)$ .

## 2 The model and the method

In the matrix variate normal setting, we aim to estimate the row and column covariance (correlation) matrices, from which we can obtain an estimate for  $\Sigma$ . The problem of covariance estimation in the context of matrix variate normal distribution is intimately connected to the problem of graphical model selection, where the graphs corresponding to the column and the row vectors are determined by the sparsity patterns (or the zeros) of  $B^{-1}$  and  $A^{-1}$  respectively. Graph estimation in this work means precisely the estimation of the zeros, as well as the non-zero entries in  $A^{-1}$  and  $B^{-1}$ . We formulate such correspondence precisely in Section 2.1. To resolve the dilemma in covariance estimation arising from unequal dimensions between the row and the column, a point which we will elaborate more in Section 3.1, we take a penalized approach to be defined in Section 2.2. We make connections between the Gemini estimators and the likelihood function of the matrix variate normal distribution in Section 2.4.

### 2.1 Problem definition: the matrix normal graphical model

We show in Figure 1 the data matrix  $X$  and its column vectors:  $x^1, x^2, \dots, x^k, \dots, x^m$ , and row vectors  $y^1, y^2, \dots, y^f$ . This notation is followed throughout the rest of the paper. First recall the following definition concerning the classical Gaussian graphical model for a random vector.

**Definition 2.1.** Let  $V = (V_1, \dots, V_f)^T$  be a random vector with distribution  $P$ , which we represent by an undirected graph  $G = (\mathcal{V}, F)$ . The vertex set  $\mathcal{V} := \{1, \dots, f\}$  has one vertex for each component of the vector  $V$ . The edge set  $F$  consists of pairs  $(j, k)$  that are joined by an edge. If  $V_j$  is independent of  $V_k$  given the other variables, then  $(j, k) \notin F$ . When  $V$  is Gaussian, missing edges correspond to zeros in the inverse covariance matrix.

Now let  $\mathcal{V} = \{1, \dots, f\}$  be an index set which enumerates rows of  $X$  according to a fixed order. For all  $i = 1, \dots, m$ , we assign to each variable of a column vector  $x^i$  exactly one element of the set  $\mathcal{V}$  by a rule of correspondence  $g : x^i \rightarrow \mathcal{V}$  such that  $g(x_j^i) = j, j = 1, \dots, f$ . The graphs  $G_i(\mathcal{V}, F)$  constructed for random column vectors  $x^i, i = 1, \dots, m$  according to Definition 2.1 will share an identical edge set  $F$ . Hence graphs  $G_1, \dots, G_m$  are isomorphic and we write  $G_i \simeq G_j, \forall i, j$ . Due to the isomorphism, we use  $G(\mathcal{V}, F)$  to represent the family of graphs  $G_1, \dots, G_m$ . Hence a pair  $(\ell, k)$  which is absent in  $F$  encodes conditional independence between the  $\ell^{th}$  row and the  $k^{th}$  row give all other rows. Similarly, let  $\Gamma = \{1, \dots, m\}$  be

the index set which enumerates columns of  $X$  according to a fixed order. We use  $H(\Gamma, E)$  to represent the family of graphs  $H_1, \dots, H_f$ , where  $H_i$  is constructed for row vector  $y^i$ , and  $H_i \simeq H_j, \forall i, j$ . Now  $H(\Gamma, E)$  is a graph with adjacency matrix  $\Upsilon(H) = \Upsilon(A^{-1})$  as edges in  $E$  encode non-zeros in  $A^{-1}$ . And  $G(\mathcal{V}, F)$  is a graph with adjacency matrix  $\Upsilon(G) = \Upsilon(B^{-1})$ . The Kronecker product,  $H \otimes G$ , is defined as the graph with adjacency matrix  $\Upsilon(H) \otimes \Upsilon(G)$  [Weichsel \(1962\)](#), where clearly missing edges correspond to zeros in the inverse covariance  $A^{-1} \otimes B^{-1}$ , and  $H \otimes G$  represents the graph of the  $p$ -variate Gaussian random vector  $\text{vec}\{X\}$ , where  $p = mf$ . In the present work, we aim to estimate  $\Upsilon(H)$  and  $\Upsilon(G)$  separately. Estimating their Kronecker product directly following the classical  $p$ -variate Gaussian graphical modeling approach will be costly in terms of both computation and the sample requirements.

## 2.2 The Gemini estimators

We start with the one-matrix case, where  $n = 1$ , and  $m$  and  $f$  are allowed to grow with respect to each other. The first hurdle we need to deal with, besides the simultaneous row and column correlations, is the fact that between the two covariance matrices  $A$  and  $B$  (as well as their inverses), the one with the higher dimension, which contains more canonical parameters, is always left with a smaller number of correlated samples in order to achieve its inference tasks. The remedy comes from the following observation. Although ambient dimension  $f, m$  can not be both bounded by the other unless  $f = m$ , the sparsity over non-diagonal entries of each precision matrix can be assumed to be bounded by the ambient dimension of the other. Under such sparsity assumptions, we first provide a pair of separable regularized estimators

$$\hat{A} = \arg \min_{A \succ 0} \left\{ \text{tr}(\hat{\Gamma}(A)A^{-1}) + \log \det A + \lambda_B \|A^{-1}\|_{1, \text{off}} \right\} \quad (2a)$$

$$\hat{B} = \arg \min_{B \succ 0} \left\{ \text{tr}(\hat{\Gamma}(B)B^{-1}) + \log \det B + \lambda_A \|B^{-1}\|_{1, \text{off}} \right\} \quad (2b)$$

for the correlation matrices  $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$  and  $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$ , where the input are a pair of sample correlation matrices  $\hat{\Gamma}(A)$  and  $\hat{\Gamma}(B)$

$$\hat{\Gamma}_{ij}(A) := \frac{\langle x^i, x^j \rangle}{\|x^i\|_2 \|x^j\|_2} \quad \text{and} \quad \hat{\Gamma}_{ij}(B) := \frac{\langle y^i, y^j \rangle}{\|y^i\|_2 \|y^j\|_2}, \quad (3)$$

and the  $\ell_1$  penalties are imposed on the off-diagonal entries of the inverse correlation estimates. Moreover, the penalty parameters  $\lambda_B$  and  $\lambda_A$ , to be defined in Section 3, are chosen to dominate the maximum of entry-wise errors for estimating  $\rho(A)$  and  $\rho(B)$  with  $\hat{\Gamma}(A)$  and  $\hat{\Gamma}(B)$  respectively. Note that the population parameters  $A$  and  $B$  can then be written as follows:

$$A \otimes B := (W_1 \rho(A) W_1) \otimes (W_2 \rho(B) W_2) / (\text{tr}(A) \text{tr}(B))$$

where  $W_1/\sqrt{\text{tr}(B)} = \text{diag}(\sqrt{a_{11}}, \dots, \sqrt{a_{mm}})$  and  $W_2/\sqrt{\text{tr}(A)} = \text{diag}(\sqrt{b_{11}}, \dots, \sqrt{b_{ff}})$ . In order to get an estimate for  $A \otimes B$ , we multiply each of the two regularized estimators  $\hat{A}$  and  $\hat{B}$  by an estimated weight matrix  $\hat{W}_1$  or  $\hat{W}_2$  respectively,

$$\hat{W}_1 = \text{diag}(\|x^1\|_2, \|x^2\|_2, \dots, \|x^m\|_2) = \text{diag}(X^T X)^{1/2}, \quad (4a)$$

$$\hat{W}_2 = \text{diag}(\|y^1\|_2, \|y^2\|_2, \dots, \|y^f\|_2) = \text{diag}(Y^T Y)^{1/2}, \quad (4b)$$

where  $Y = X^T$ . Up to a multiplicative factor  $\text{tr}(B)$  and  $\text{tr}(A)$ ,  $\widehat{W}_1^2$  and  $\widehat{W}_2^2$  will provide an estimate for  $\text{diag}(A)$  and  $\text{diag}(B)$  respectively; To estimate  $A \otimes B$ , we compute the Kronecker product of our weighted estimators,

$$\widehat{A \otimes B} := (\widehat{W}_1 \widehat{A} \widehat{W}_1) \otimes (\widehat{W}_2 \widehat{B} \widehat{W}_2) / \|X\|_F^2$$

while adjusting the unknown multiplicative factors  $\text{tr}(B)\text{tr}(A)$  by  $\|X\|_F^2$ . We restate Theorem 4.1 in case  $n = 1$  in Corollary 2.2.

**Corollary 2.2.** *Let  $\max(m, f) \geq 2$ . Then with probability at least  $1 - \frac{3}{\max(m, f)^2}$ ,  $\forall i \neq j$ , we have for  $\widehat{\Gamma}(A)$  and  $\widehat{\Gamma}(B)$  are constructed as in (3),*

$$\begin{aligned} \left| \widehat{\Gamma}_{ij}(B) - \rho_{ij}(B) \right| &= \left| \frac{\langle y^i, y^j \rangle}{\|y^i\|_2 \|y^j\|_2} - \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}} \right| \leq \frac{\alpha}{1 - \alpha} + |\rho_{ij}(B)| \frac{\alpha}{1 - \alpha}, \\ \left| \widehat{\Gamma}_{ij}(A) - \rho_{ij}(A) \right| &= \left| \frac{\langle x^i, x^j \rangle}{\|x^i\|_2 \|x^j\|_2} - \frac{a_{ij}}{\sqrt{a_{ii} a_{jj}}} \right| \leq \frac{\beta}{1 - \beta} + |\rho_{ij}(A)| \frac{\beta}{1 - \beta} \end{aligned}$$

and  $\left| \|X\|_F^2 / (\text{tr}(A)\text{tr}(B)) - 1 \right| \leq \alpha \wedge \beta$ , where

$$\alpha = 20 \frac{\|A\|_F \log \max(m, f)}{\text{tr}(A)}, \quad \text{and} \quad \beta = 20 \frac{\|B\|_F \log \max(m, f)}{\text{tr}(B)}.$$

### 2.3 Gemini for replicates of $X$

We now adapt the Gemini estimators as defined in Section 2.2 to the general setting where we have multiple replicates of  $X$ . Suppose that we have  $n$  independently and identically distributed matrices  $X(1), \dots, X(n) \sim \mathcal{N}_{f,m}(0, A \otimes B)$ . For each  $t$ , we denote by

$$X(t) = [x(t)^1 x(t)^2 \dots x(t)^m] = [y(t)^1 y(t)^2 \dots y(t)^f]^T \quad (5)$$

the matrix  $X_{f \times m}(t)$  with  $x(t)^1, \dots, x(t)^m \in \mathbf{R}^f$  being its columns vectors and  $y^1(t), \dots, y^f(t)$  being its row vectors.

First, we update our sample correlation matrices:

$$\widehat{\Gamma}_{ij}(A) := \frac{\sum_{t=1}^n \langle x(t)^i, x(t)^j \rangle}{\sqrt{\sum_{t=1}^n \|x(t)^i\|_2^2} \sqrt{\sum_{t=1}^n \|x(t)^j\|_2^2}}, \quad (6a)$$

$$\widehat{\Gamma}_{ij}(B) := \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{\sum_{t=1}^n \|y(t)^i\|_2^2} \sqrt{\sum_{t=1}^n \|y(t)^j\|_2^2}}. \quad (6b)$$

We will show in Theorem 4.1 large deviation bounds for estimating the correlation coefficients in  $\rho(A)$  and  $\rho(B)$  with entries in sample correlation  $\widehat{\Gamma}(A)$  and  $\widehat{\Gamma}(B)$  constructed above. Plugging these updated  $\widehat{\Gamma}(A)$  and  $\widehat{\Gamma}(B)$  in (2a) and (2b) will return us the penalized correlation estimators  $\widehat{A}$  and  $\widehat{B}$ . Next, we update the weight matrices  $\widehat{W}_1$  and  $\widehat{W}_2$  as follows:

$$\widehat{W}_1 = \text{diag} \left( \sqrt{\frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2}, i = 1, \dots, m \right), \quad (7a)$$

$$\widehat{W}_2 = \text{diag} \left( \sqrt{\frac{1}{n} \sum_{t=1}^n \|y(t)^j\|_2^2}, j = 1, \dots, f \right). \quad (7b)$$

We can then construct an estimator for  $A \otimes B$  as before,

$$\widehat{A \otimes B} := \left( \widehat{W_1} \widehat{A} \widehat{W_1} \right) \otimes \left( \widehat{W_2} \widehat{B} \widehat{W_2} \right) / \left( \frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 \right). \quad (8)$$

In Section 3, we state the convergence rates for estimating  $A \otimes B$  and its inverse with  $\widehat{A \otimes B}$  and  $\widehat{A \otimes B}^{-1}$ .

## 2.4 Discussion

As mentioned, these optimization goals in (2a) and (2b) are intimately connected to the likelihood function of the matrix variate normal distribution. Let  $X^n = (X(1), \dots, X(n))$ , where  $X(i) \sim \mathcal{N}_{f,m}(M, A \otimes B)$ , we have up to an additive constant,

$$-2 \log p(X^n | M, A, B) = \text{tr} \left( (A^{-1} \otimes B^{-1}) \widehat{S}_n \right) + f \log |A| + m \log |B|$$

where  $\widehat{S}_n = \frac{1}{n} \sum_{i=1}^n \text{vec} \{ X(t) - M \} \text{vec} \{ X(t) - M \}^T$  is the sample covariance matrix for  $\text{vec} \{ X \}$ . We assume that  $M = 0$ . There is no known closed-form solution for the maximum of the likelihood function for the matrix-variate normal distribution. Essentially the flip-flop methods Dutilleul (1999); Lu and Zimmerman (2005) couple the estimation for  $A$  and  $B$  by feeding the current estimate for either of the two into the likelihood function (or the penalized variants to be defined) in order to optimize it with respect to the other. Upon initialization of  $A$  with an identity matrix, they obtain the MLE for  $A$  and  $B$  by solving the following two equations alternately and iteratively

$$\widetilde{B}(A) = \frac{1}{nm} \sum_{t=1}^n X(t) A^{-1} X(t)^T, \quad \widetilde{A}(B) = \frac{1}{nf} \sum_{t=1}^n X(t)^T B^{-1} X(t) \quad (9)$$

such that the corresponding output  $\widetilde{B}$ , or  $\widetilde{A}$  becomes the input as  $B$ , or  $A$  to the RHS of equations in (9); this process repeats until certain convergence criteria are reached.

In the present work, we simultaneously optimize a pair of convex functions (2a) and (2b), which can be seen as a single-step approximation of a penalized version of (9), where we set both  $B$  and  $A$  on the RHS of equations in (9) to be the identity matrix. To see this, we note that

$$\begin{aligned} \widehat{\Gamma}(A) &= \text{diag}(\widetilde{A}(I))^{-1/2} \widetilde{A}(I) \text{diag}(\widetilde{A}(I))^{-1/2} \\ \text{and } \widehat{\Gamma}(B) &= \text{diag}(\widetilde{B}(I))^{-1/2} \widetilde{B}(I) \text{diag}(\widetilde{B}(I))^{-1/2} \end{aligned}$$

where  $\widehat{\Gamma}(A), \widehat{\Gamma}(B)$  are the sample correlation matrices as defined in (6a) and (6b),  $\widetilde{A}(I) = \frac{1}{n} \sum_{t=1}^n \frac{1}{f} \sum_{k=1}^f y(t)^k \otimes y(t)^k$  are the average of diagonal blocks in  $\widetilde{S}_n$ , which is the sample covariance matrix for  $\text{vec} \{ X^T \}$ , and  $\widetilde{B}(I) = \frac{1}{n} \sum_{t=1}^n \frac{1}{m} \sum_{k=1}^m x(t)^k \otimes x(t)^k$  are the average of those in  $\widehat{S}_n$ .

Closest to our methods is that of Allen and Tibshirani (2010). They aim to optimize  $\mathcal{L}(A, B) := \text{tr}((A^{-1} \otimes B^{-1}) \widehat{S}_n) + m \log \det B + f \log \det A + \lambda_1 |A^{-1}|_1 + \lambda_2 |B^{-1}|_1$  over both  $A$  and  $B$ , where  $|\cdot|_1$  penalty can also be replaced with the Frobenius norm penalty instead. They argue that using two separate penalty parameters or even two kinds of penalties gives greater flexibility in terms of separately modeling the row and column covariances. A popular method for minimizing the objective  $\mathcal{L}(A, B)$  is the block coordinate descent method Tseng (2001). Following essentially the arguments in Tseng (2001), they show that while



block coordinate-wise maximization reaches a stationary point, it is not guaranteed to reach the global maximum of the bi-convex optimization function of  $A^{-1}$  and  $B^{-1}$ .

In contrast, in Gemini, a unique pair of optimal solutions  $\hat{A}$  and  $\hat{B}$  can be obtained via the GLasso algorithm: Upon computing the sample correlation matrices  $\hat{\Gamma}(A)$  and  $\hat{\Gamma}(B)$  as in (6a) and (6b), one can then rely on publicly available software such as the `glasso` package in R, which implements the graphical Lasso algorithm [Friedman et al. \(2008\)](#) to solve (2a) and (2b). The advantage of estimating the graphs for  $A^{-1}$  and  $B^{-1}$  via separable optimization functions is motivated by both statistical and computational considerations. Under sparsity constraints and upon multiplication by proper weight matrices, the penalized estimators  $\hat{A}$  and  $\hat{B}$  are strikingly effective in approximating the row and column covariance matrices. See Theorem 4.3 and Corollary A.3. Our approach is computational efficient, in that, the main cost of our estimators involves solving two GLasso problems, the total computational cost of which is in the order of  $O(f^3 + m^3)$  for sparse graphs or  $O(f^4 + m^4)$  for general graphs, while solving a global GLasso program in general will cost between  $O(f^3 m^3)$  and  $O(f^4 m^4)$ . This will become prohibitive for  $f, m$  in the order of 100.

We use theoretical analysis to illustrate the advantage of choosing separate penalty parameters on the estimated inverse correlations and covariances. In practice, we use cross-validation to select the penalty parameters as we will show in our numerical examples. It is conceivable that the iterative methods are much more computational demanding in that it requires cross-validation at each iteration. Hence we study a non-iterative penalized flip-flop algorithm and present its convergence properties in Section 5 and 6.

### 3 Theoretical results

In this section, we present in Theorem 3.1 and Theorem 3.2 the convergence rates for estimating the row and column covariance matrices and their inverses with respect to the operator norm and the Frobenius norm respectively. Our analysis is non-asymptotic in nature; however, we first formulate our results from an asymptotic point of view for simplicity. To do so, we consider an array of matrix variate normal data

$$X(1), \dots, X(n) \text{ i.i.d. } \sim \mathcal{N}_{f,m}(0, A_0 \otimes B_0), \quad n = 1, 2, \dots \quad (10)$$

where  $f, m, |A_0^{-1}|_{0,\text{off}}$ , and  $|B_0^{-1}|_{0,\text{off}}$ , which are the number of non-zero non-diagonal entries in the inverse covariance matrices, may change with  $n$ .

We make the following assumptions.

- (A1) The dimensions of  $f, m$  are allowed to grow with respect to each other and for  $d = 1 - (c/2 \wedge 1/2)$  where  $c = \frac{\log n}{\log \log(m \vee f)}$ ,

$$\begin{aligned} |A_0^{-1}|_{0,\text{off}} &= o\left(nf / \log^{2d}(m \vee f)\right) \quad (f, m \rightarrow \infty), \\ \text{and } |B_0^{-1}|_{0,\text{off}} &= o\left(nm / \log^{2d}(m \vee f)\right) \quad (f, m \rightarrow \infty). \end{aligned}$$

- (A2) The eigenvalues  $\varphi_i(A_0), \varphi_j(B_0), \forall i, j$  of the positive definite covariance matrices  $A_0$  and  $B_0$  are bounded away from 0 and  $+\infty$ .

For  $n = 1$ , (A1) implies that, up to a logarithmic factor, the number of non-zero off-diagonal entries in  $\rho(A_0)^{-1}$  or  $\rho(B_0)^{-1}$  must be bounded by the dimension of the other matrix. Let  $a_{\min} = \min_i A_{0,ii}$  and



$b_{\min} = \min_i B_{0,ii}$ . By positive-definiteness of  $A_0$  and  $B_0$ , we have  $a_{\min} > 0$  and  $b_{\min} > 0$ . Then clearly for  $a_{\max} = \max_i A_{0,ii}$  and  $b_{\max} = \max_i B_{0,ii}$ , we have  $0 < a_{\min} \leq a_{\max} \leq \|A_0\|_2 < +\infty$  and  $0 < b_{\min} \leq b_{\max} \leq \|B_0\|_2 < +\infty$  on (A2).

We now state the main results of this paper. These results are new to the best of our knowledge. These bounds are stated in terms of the relative errors; The absolute error bounds are given in Theorem A.1 and Theorem A.2. The proofs for Theorems 3.1 and 3.2 appear in Sections A.2 and A.3.

**Theorem 3.1.** Consider data generating random matrices as in (10) and assume that (A1) and (A2) hold. Let  $c = \frac{\log n}{\log \log(m \vee f)}$  and  $d = 1 - (c/2 \wedge 1/2)$ . Suppose the penalty parameters are chosen to be

$$\begin{aligned} \lambda_A = \lambda_{A_0} &\asymp \frac{\|A_0\|_F \log^d \max(m, f)}{\text{tr}(A_0) \sqrt{n}} \asymp \frac{\log^d \max(m, f)}{\sqrt{mn}} \rightarrow 0, \\ \text{and } \lambda_B = \lambda_{B_0} &\asymp \frac{\|B_0\|_F \log^d \max(m, f)}{\text{tr}(B_0) \sqrt{n}} \asymp \frac{\log^d \max(m, f)}{\sqrt{fn}} \rightarrow 0. \end{aligned}$$

Then with probability at least  $1 - \frac{3}{\max(m, f)^2}$ , for  $\widehat{A \otimes B}$  as defined in (8),

$$\begin{aligned} \left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_2 &\leq \|A_0\|_2 \|B_0\|_2 \delta, \text{ and} \\ \left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\|_2 &\leq \|B_0^{-1}\|_2 \|A_0^{-1}\|_2 \delta', \text{ where} \\ \delta, \delta' &\asymp \log^d \max(m, f) \left( \sqrt{\frac{|A_0^{-1}|_{0, \text{off}} \vee 1}{nf}} + \sqrt{\frac{|B_0^{-1}|_{0, \text{off}} \vee 1}{nm}} \right) = o(1). \end{aligned}$$

**Theorem 3.2.** Consider data generating random matrices as in (10). Let  $\lambda_{A_0}$  and  $\lambda_{B_0}$  be chosen as in Theorem 3.1. Let  $\widehat{A \otimes B}$  be as defined in (8). Under Assumptions (A1) and (A2), we have

$$\begin{aligned} \left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F &\leq \delta \|A_0\|_F \|B_0\|_F \text{ where} \\ \delta &= O \left( \lambda_{B_0} \sqrt{|A_0^{-1}|_{0, \text{off}} \vee m / \sqrt{m}} + \lambda_{A_0} \sqrt{|B_0^{-1}|_{0, \text{off}} \vee f / \sqrt{f}} \right) = o(1). \end{aligned} \quad (11)$$

In particular, suppose (1)  $1 \leq n \leq \log(m \vee f)$  Or (2)  $|A_0^{-1}|_{0, \text{off}} = O(m)$  and  $|B_0^{-1}|_{0, \text{off}} = O(f)$ . Then under (A1) and (A2), we have  $\delta \asymp \lambda_{A_0} + \lambda_{B_0}$  and

$$\left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F \leq \delta \|A_0\|_F \|B_0\|_F \asymp \frac{\log^d(m \vee f)}{\sqrt{n}} (\sqrt{m} + \sqrt{f}). \quad (12)$$

The same conclusions hold for the inverse estimate

$$\left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\|_F \leq \delta \|A_0^{-1}\|_F \|B_0^{-1}\|_F \quad (13)$$

with  $\delta$  being bounded in the same order as above for each case.

**Remark 3.3.** Let us examine the rate of (11) for some other cases. Suppose that  $|A_0^{-1}|_{0, \text{off}} \geq m$  and  $|B_0^{-1}|_{0, \text{off}} \geq f$  and  $n > \log(m \vee f)$  is moderately large. Then under (A1)

$$\begin{aligned} \delta &= O \left( \lambda_{B_0} \sqrt{|A_0^{-1}|_{0, \text{off}} / m} + \lambda_{A_0} \sqrt{|B_0^{-1}|_{0, \text{off}} / f} \right) = o \left( \frac{1}{\sqrt{f}} + \frac{1}{\sqrt{m}} \right) \\ \text{and } \left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F &\leq \delta \|A_0\|_F \|B_0\|_F \asymp o \left( \sqrt{m} + \sqrt{f} \right) \end{aligned} \quad (14)$$

where we assume that both  $|A_0^{-1}|_{0,\text{off}}$  and  $|B_0^{-1}|_{0,\text{off}}$  are allowed to grow linearly with  $n$  in the worst case, and hence the bound on  $\delta$  becomes independent of  $n$ . We note that for the rate we calculated in (14), we assume that  $n$  is not too large; otherwise, one can refine the calculations a bit further.

For example, suppose that  $n > \log \max(m, f) \left( \frac{f^2}{m} \vee \frac{m^2}{f} \right)$ . Then (A1) becomes vacuous and trivially, we have for  $d = 1/2$ ,

$$\delta = O\left(\lambda_{B_0}\sqrt{m} + \lambda_{A_0}\sqrt{f}\right) = O\left(\frac{m + f \log^{1/2} \max(m, f)}{\sqrt{fm}} \frac{1}{\sqrt{n}}\right).$$

Hence when  $n \asymp \log \max(m, f) \left( \frac{f^2}{m} \vee \frac{m^2}{f} \right)$ , we have  $\delta = o\left(\frac{1}{\sqrt{m}} \wedge \frac{1}{\sqrt{f}}\right)$  and

$$\left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F \leq \delta \|A_0\|_F \|B_0\|_F \asymp o(\sqrt{f} \wedge \sqrt{m})$$

We do not pursue such refinements in this work.

### 3.1 Discussion

To put our discussions on the rates of convergence for covariance estimation in context, we first present an example from the classical multivariate analysis. Consider the case where we are given a single sample from the matrix variate normal distribution with  $B_0 = I$ , and the dimensions  $f, m$  increase to infinity, while the aspect ratio  $f/m \rightarrow \text{const} > 1$ .

The classical multivariate analysis focuses on estimating  $A_0$  using data matrix  $X$  whose rows are independent replicates following  $\mathcal{N}_m(0, A_0)$ . The simplest way to estimate  $A_0$  is to compute the sample covariance

$$\tilde{A}_f = \frac{1}{f} X^T X = \frac{1}{f} \sum_{i=1}^f x_i \otimes x_i, \text{ where } x_i \sim \mathcal{N}_m(0, A_0).$$

The problem here is to determine the minimal number of independent rows we need so that the sample covariance matrix  $\tilde{A}_f$  approximates  $A$  “well” in the operator norm. This concerns the classical “Bai-Yin law” in random matrix theory regarding the Wishart random matrix  $\tilde{A}_f = \frac{1}{f} X^T X$ , which says that the spectrum of  $\tilde{A}_f$  is almost surely contained in the interval  $[a^2/f + o(1), b^2/f + o(1)]$  where  $a = (\sqrt{f} - \sqrt{m})_+$  and  $b = \sqrt{f} + \sqrt{m}$  in case  $A_0 = I$ . For general covariance matrix  $A_0$ , it follows that,

$$\|\tilde{A}_f - A_0\|_2 \leq \left(2\sqrt{m/f} + (m/f) + o(1)\right) \|A_0\|_2 \quad (15)$$

with high probability. We refer to [Vershynin \(2012\)](#) for a comprehensive exposition on such results. While such results provide a satisfactory answer to the covariance estimation problem in the regime  $f \geq m$  for general multivariate normal distributions, it remains challenging problems as to: (1) how to estimate the covariance matrix which has the larger dimension of the two? that is, how can we approximate  $A_0$  well in the operator norm when  $f < m$ ? (2) how to estimate both  $A_0$  and  $B_0$  given both correlated rows and columns?

Our answer to the first question is to use the penalized methods. The operator norm bound in Theorem 3.1 illustrates the point that the combination of sparsity and spectral assumptions as in (A1), (A2), and  $\ell_1$ -regularization ensures convergence on estimation of the covariance and precision matrices, even though

their ambient dimensions may greatly exceed the given sample sizes. In particular, the ambient dimensions which appear in the numerator in (15) are replaced with the sparsity parameters:

$$\delta, \delta' \asymp \log \max(m, f) \left( \frac{1}{\sqrt{f}} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} + \frac{1}{\sqrt{m}} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} \right) = o(1).$$

which holds for  $n = 1$  with high probability under A(1) and (A2). We also pay an extra logarithmic factor for such dependences between both rows and columns. We note that between  $A_0$  and  $B_0$ , the dimension of one matrix is the same as the number of samples available for estimating parameters in the other matrix. For example, the number of examples we have for estimating parameters in matrix  $A_0$  using the data matrix  $X$  as shown in Figure 1 is  $f$ , which is the dimension of matrix  $B_0$ ; and vice versa. The two summands in  $\delta$  and  $\delta'$  in Theorems 3.1 reflect the rates of convergence in the Frobenius norm for estimating  $\rho(A_0)$  and  $\rho(B_0)$  with  $\hat{A}$  and  $\hat{B}$  as in (2a) and (2b) (cf. Theorem 4.3). Moreover, the two summands in  $\delta$  and  $\delta'$  in Theorems 3.1 and 3.2 also correspond to the rates of convergence in the operator and the Frobenius norm for estimating the row and column covariance matrices  $A_0, B_0$ , up to a scale factor, respectively. In particular, Corollary A.3 provides rates of convergence for estimating  $A_*, B_*$  (and their inverses), where

$$A_* = mA_0/\text{tr}(A_0) \quad \text{and} \quad B_* = B_0\text{tr}(A_0)/m, \quad (16)$$

in both norms with the following estimators:

$$\hat{A}_* = m\widehat{W}_1\widehat{A}\widehat{W}_1/(\frac{1}{n}\sum_{i=1}^n\|X(i)\|_F^2) \quad \text{and} \quad \hat{B}_* = \widehat{W}_2\widehat{B}\widehat{W}_2/m \quad (17)$$

where clearly  $\hat{A}_* \otimes \hat{B}_* = \widehat{A \otimes B}$ .

To answer the second question, first note that

$$\begin{aligned} \mathbb{E}[X^T X/f] &= \mathbb{E}\left[\frac{1}{f}\sum_{i=1}^f y^i \otimes y^i\right] = A_0\text{tr}(B_0)/f \\ \mathbb{E}[X X^T/m] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m x^i \otimes x^i\right] = B_0\text{tr}(A_0)/m, \end{aligned}$$

which suggest that  $\tilde{A}_f$  and  $\tilde{B}_m = X X^T/m$  are good starting points for us to construct estimators for  $\rho(A_0)$ ,  $\rho(B_0)$ , and their inverses, despite the presence of dependence along the other dimension. We construct more sophisticated sample covariance and correlation estimators based on the pair of likelihood functions in (9) in Section 5. The relationships between the row correlations and column correlations of  $X$  are known to complicate the solution to the related problem of testing the hypothesis that microarrays are independent of each other given possibly correlated genes. For example, it was observed in Efron (2009) that the gene-wise correlation can induce the appearance of microarray-wise correlation through the “leakage” phenomenon in the doubly standardized situation for  $X$ . We illustrate such interactions between the row-wise and column-wise correlations and inverse correlations via the large deviation bounds which we derive in Section 6.

To understand the rates for  $n > 1$  and to build the connection between the one-matrix and the multiple-matrix cases, we imagine stacking matrices  $X(1), \dots, X(n)$  on top of each other to form a single  $nf \times m$  matrix  $X'$ . This way, we can then imagine being in the situation of one-matrix case:  $X' \sim \mathcal{N}_{nf,m}(0, A_0 \otimes B')$ , where the dimension of  $A_0$  and  $B'$  is  $m \times m$  and  $nf \times nf$  respectively. Similar to the general case with one data matrix, we will use a sample of size  $nf$  to estimate the structure and parameters for  $A_0$ ; however, unlike the one matrix case we have focused on so far,  $B'$  has an additional structural property other than the assumed sparsity on its inverse which allows for a faster rate of convergence, namely, matrix  $B'$  is block diagonal with  $n$  identical submatrices  $B_0$  along its diagonal. Hence the estimation step for  $B'$

takes advantage of this knowledge by stacking  $Y(1), \dots, Y(n)$  on top of each other, where  $Y(i) = X(i)^T$ , to form a  $nm \times f$  matrix  $Y' \sim \mathcal{N}_{nm,f}(0, B_0 \otimes A')$ , where matrix  $A'$  is in turn block diagonal with  $n$  identical submatrices  $A_0$  staying along its diagonal. We treat  $Y'$  as having  $nm$  samples, all of which subject to normalization come from the same multivariate normal distribution  $\mathcal{N}_f(0, B_0)$  as shown in Figure 1. This enables faster convergence rates for  $n > 1$  as shown in Theorems 3.1 and 3.2.

Finally, in our model, we assume that both the rows and columns of  $X$  are centered by setting the mean matrix  $M = 0$ , but we do not assume that they are also scaled. That is, we allow the diagonals of  $A$  and  $B$  to take arbitrary positive real numbers which are bounded away from 0 or  $+\infty$  as assumed in (A2). It would be interesting to study the following mean-restricted matrix-variate normal model:  $X_{ij} = \nu_i + \mu_j + \varepsilon_{ij}$ , where the two additive fixed effects  $\nu_i, \mu_j$  depend on the row and column means and  $\varepsilon \sim \mathcal{N}_{f,m}(0, A_0 \otimes B_0)$  is a mean-zero random effect. See Bonilla et al. (2008); Yu et al. (2009); Allen and Tibshirani (2010) and references therein for applications of such and other random effects models.

### 3.2 Related work

In the classical setting, various work Dutilleul (1999); Lu and Zimmerman (2005); Werner et al. (2008) focused on algorithms and convergence properties on estimating  $\Sigma$  using a large number of samples  $X(1), \dots, X(n)$ . The flip-flop methods for estimating  $A_0 \otimes B_0$  with a finite number of iterations have been shown to converge asymptotically as well as the MLE as  $n \rightarrow \infty$  Werner et al. (2008). Other recent work with an iterative approach for solving the graphical model selection problem in the context of matrix variate normal distribution include Zhang and Schneider (2010); Yin and Li (2012); Tsiglikaridis et al. (2012). None of these work was able to show convergence in the operator norm which works in case  $n = 1$  and  $f, m \rightarrow \infty$  as in our work. When  $f, m$  diverge as  $n \rightarrow \infty$ , the rates in Yin and Li (2012) are significantly slower than the corresponding ones in the present work. In particular, certain non-asymptotic rates in the Frobenius norm are obtained in estimating  $A_0$  and  $B_0$ ; However, while estimating  $B_0$  or  $A_0$ , the ambient dimension of the other matrix, namely,  $A_0$  or  $B_0$  appears in the numerator instead of the denominator. This is undesirable, as we have discussed in Section 3.1 that the effective sample size in estimating  $B_0$  and  $A_0$  should be  $nm$  and  $nf$  respectively. Following essentially the same methods as in Allen and Tibshirani (2010), the same convergence rate as in (12) on estimating the covariance  $\Sigma = A_0 \otimes B_0$  in the Frobenius norm is obtained in Tsiglikaridis et al. (2012), upon a finite number of iterations in case  $|A_0^{-1}|_{0,\text{off}} = O(m)$ , and  $|B_0^{-1}|_{0,\text{off}} = O(f)$ ; however, this rate is obtained with the additional requirement that the number of replicates of  $X$  must be at least the order of  $n \geq \Omega\left(\left(\frac{f}{m} \vee \frac{m}{f}\right) \log \max(f, m, n)\right)$ . This excludes the case for  $n = 1$  or for  $n < \log(m \vee f)$ , which is the main focus of the present paper.

High-dimensional covariance (inverse) estimation has been intensively studied in the literature under various structural assumptions, see for example Huang et al. (2006); Furrer and Bengtsson (2007); Meinshausen and Bühlmann (2006); Lam and Fan (2009); Bickel and Levina (2008); El Karoui (2008); Yuan and Lin (2007); Friedman et al. (2008); Banerjee et al. (2008); Rothman et al. (2008); Ravikumar et al. (2008); Peng et al. (2009); Yuan (2010); Cai et al. (2010, 2011); Zhou et al. (2011). It is plausible that we can extend some of these methods and techniques on sparse (inverse) covariance estimation and apply to the matrix variate setting under a different set of assumptions. A nonparametric method for estimating time varying graphical structures using independent but not identical samples with provably convergence rates under suitable assumptions was studied by Zhou et al. (2008). It would be interesting to consider such models in the matrix variate normal setting.

## 4 Estimation of the correlation coefficients

In this section, we elaborate on two key technical results, namely, the concentration bounds for sample correlation estimates and the convergence bounds for the penalized inverse correlation estimates. We use these results to prove the main theorems 3.1 and 3.2.

### 4.1 Concentration bounds for sample correlations

We now show the concentration bounds for estimating the parameters in  $\rho(A_0)$  and  $\rho(B_0)$ . Theorem 4.1 covers the “finite” sample setting where the number of replications  $n$  being upper bounded by  $\log(m \vee f)$ . We believe these are the first of such results to the best of our knowledge. For completeness, we also state the bounds when  $n$  is large. The proof appears in Section B.2.

**Theorem 4.1.** *Let  $1 \leq n \leq \log \max(m, f)$ . Consider data generating random matrices as in (10). Let us define for  $c = \frac{\log n}{\log \log(m \vee f)}$ , where  $\max(m, f) \geq 2$ , and  $d = 1 - c/2$ . Let*

$$\begin{aligned}\alpha_n &:= m\tau_0/\text{tr}(A_0) \text{ where } \tau_0 = 20(\|A_0\|_F/\sqrt{m})\log^d \max(m, f)/\sqrt{mn}, \\ \beta_n &:= f\tau'_0/\text{tr}(B_0) \text{ where } \tau'_0 = 20(\|B_0\|_F/\sqrt{f})\log^d \max(m, f)/\sqrt{fn}.\end{aligned}$$

*Then with probability at least  $1 - \frac{3}{\max(m, f)^2}$ , for  $\alpha_n, \beta_n < 1/3$ , and  $\hat{\Gamma}(A_0)$  and  $\hat{\Gamma}(B_0)$  as in (6a) and (6b), we have*

$$\begin{aligned}\forall i \neq j, \quad & \left| \hat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \right| \leq \frac{\alpha_n}{1 - \alpha_n} + |\rho_{ij}(B_0)| \frac{\alpha_n}{1 - \alpha_n} \leq 3\alpha_n, \\ \forall i \neq j, \quad & \left| \hat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \right| \leq \frac{\beta_n}{1 - \beta_n} + |\rho_{ij}(A_0)| \frac{\beta_n}{1 - \beta_n} \leq 3\beta_n, \\ \text{and} \quad & \left| \frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 - \text{tr}(A_0)\text{tr}(B_0) \right| \leq \text{tr}(A_0)\text{tr}(B_0)(\alpha_n \wedge \beta_n).\end{aligned}\tag{18}$$

*When  $n > \log(m \vee f)$ , we always set  $d = 1/2$  in  $\tau_0$  and  $\tau'_0$ , where 20 is replaced with a smaller constant to be defined in Theorem C.1. Then all three inequalities immediately above will continue to hold.*

The following large deviation bounds in Lemma 4.2 are the key results in proving Theorem 4.1. We write it explicitly to denote by  $\mathcal{X}_0$  the event that all large deviation inequalities as stated in Lemma 4.2 hold. To establish such concentration bounds, we apply decorrelation techniques and the factorial-type moment estimate for the Gaussian chaos from Ledoux and Talagrand (1991) to  $\langle x^i, x^j \rangle$ , and  $\langle y^i, y^j \rangle$ , for each  $i, j$ , which make applications of Bernstein’s and Bennet’s inequalities Bennett (1962) possible. We leave the theoretical formulation and proofs in Sections B and B.3.

**Lemma 4.2.** *Denote by  $\mathcal{X}_0$  the event that the following inequalities hold simultaneously for  $\tau_0, \tau'_0$  as defined in Theorem 4.1:*

$$\begin{aligned}\forall i, \quad & \frac{1}{m} \left| \frac{1}{n} \sum_{t=1}^n (\|y(t)^i\|_2^2 / b_{ii}) - \text{tr}(A_0) \right| \leq \tau_0, \\ \forall i \neq j, \quad & \frac{1}{m} \left| \frac{1}{n} \sum_{t=1}^n (\langle y(t)^i, y(t)^j \rangle / \sqrt{b_{ii}b_{jj}}) - \rho_{ij}(B_0)\text{tr}(A_0) \right| \leq \tau_0, \\ \forall i, \quad & \frac{1}{f} \left| \frac{1}{n} \sum_{t=1}^n (\|x(t)^i\|_2^2 / a_{ii}) - \text{tr}(B_0) \right| \leq \tau'_0, \\ \forall i, j \quad & \frac{1}{f} \left| \frac{1}{n} \sum_{t=1}^n (\langle x(t)^i, x(t)^j \rangle / \sqrt{a_{ii}a_{jj}}) - \rho_{ij}(A_0)\text{tr}(B_0) \right| \leq \tau'_0.\end{aligned}$$

*Then  $\mathbb{P}(\mathcal{X}_0) \geq 1 - \frac{3}{\max(m, f)^2}$ .*

## 4.2 Bounds on estimating the inverse correlation matrices

In this section, we show explicit non-asymptotic convergence rates in the Frobenius norm for estimating  $\rho(A_0)$ ,  $\rho(B_0)$ , and their inverses in Theorem 4.3.

We say that event  $\mathcal{T}(A_0)$  holds for sample correlation matrix  $\hat{\Gamma}(A_0)$  for some parameter  $\delta_{n,f} \rightarrow 0$ , if for all  $j$ ,  $\hat{\Gamma}_{jj}(A_0) = \rho_{jj}(A_0) = 1$  and

$$\max_{j,k,j \neq k} |\hat{\Gamma}_{jk}(A_0) - \rho_{jk}(A_0)| \leq \delta_{n,f}, \quad (19)$$

and the event  $\mathcal{T}(B_0)$  holds for sample correlation matrix  $\hat{\Gamma}(B_0)$  for some parameter  $\delta_{n,m} \rightarrow 0$ , if for all  $j$ ,  $\hat{\Gamma}_{jj}(B_0) = \rho_{jj}(B_0) = 1$  and

$$\max_{j,k,j \neq k} |\hat{\Gamma}_{jk}(B_0) - \rho_{jk}(B_0)| \leq \delta_{n,m}. \quad (20)$$

**Theorem 4.3.** *Suppose that (A1) and (A2) hold. Let  $\hat{A}$  and  $\hat{B}$  be the unique minimizers defined by (2a) and (2a) with sample correlation matrices  $\hat{\Gamma}(A_0)$  and  $\hat{\Gamma}(B_0)$  as their input. Suppose that event  $\mathcal{T}(A_0)$  holds for  $\hat{\Gamma}(A_0)$  for some  $\delta_{n,f}$  and event  $\mathcal{T}(B_0)$  holds for  $\hat{\Gamma}(B_0)$  for some  $\delta_{n,m}$ , such that*

$$\begin{aligned} \delta_{n,f} \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1 &= o(1) \quad \text{and} \quad \delta_{n,m} \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1 = o(1). \\ \text{Set for some } 0 < \epsilon, \epsilon < 1, \quad \lambda_B &= \delta_{n,f}/\epsilon \quad \text{and} \quad \lambda_A = \delta_{n,m}/\epsilon. \end{aligned} \quad (21)$$

Then on event  $\mathcal{T}(A_0) \cap \mathcal{T}(B_0)$ , we have for  $C = 9(1 + \epsilon)/2$  and  $C' = 9(1 + \epsilon)/2$ ,

$$\left\| \hat{A}^{-1} - \rho(A_0)^{-1} \right\|_F < C \lambda_B \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1 / \varphi_{\min}^2(\rho(A_0)), \quad (22)$$

$$\left\| \hat{B}^{-1} - \rho(B_0)^{-1} \right\|_F \leq C' \lambda_A \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1 / \varphi_{\min}^2(\rho(B_0)), \quad (23)$$

$$\begin{aligned} \left\| \hat{A} - \rho(A_0) \right\|_F &\leq 2C \kappa(\rho(A_0))^2 \lambda_B \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1, \\ \text{and} \quad \left\| \hat{B} - \rho(B_0) \right\|_F &\leq 2C' \kappa(\rho(B_0))^2 \lambda_A \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1. \end{aligned}$$

Corollary 4.4 provides a bound on the off-diagonal  $\ell_1$  norm on the error matrices for estimating  $\Theta_0 = \rho(A_0)^{-1}$  and  $\Phi_0 = \rho(B_0)^{-1}$ , which may be of independent interests. We use it in our analysis of the flip-flip algorithm.

**Corollary 4.4.** *Suppose conditions in Theorem 4.3 hold, and  $\lambda_B$  and  $\lambda_A$  are chosen as in (21). Let  $S = \{(i, j) : \Theta_{0ij} \neq 0, i \neq j\}$  and  $S^c = \{(i, j) : \Theta_{0ij} = 0, i \neq j\}$ . Then on  $\mathcal{T}(A_0)$  as defined in Theorem 4.3,*

$$|\Delta_{S^c}|_1 \leq \frac{1 + \epsilon}{1 - \epsilon} |\Delta_S|_1 \quad \text{where} \quad \Delta = \hat{A}^{-1} - \Theta_0 =: \Delta_{A_0} \quad (24)$$

$$\text{and} \quad |\Delta_{A_0}|_{1,\text{off}} \leq \frac{1 + \epsilon}{1 - \epsilon} 9 \lambda_B \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1 \sqrt{|A_0^{-1}|_{0,\text{off}}} / \varphi_{\min}^2(\rho(A_0)).$$

Similarly, for  $\Delta := \Delta_{B_0} = \hat{B}^{-1} - \Phi_0$ ,  $|\Delta_{S^c}|_1 \leq \frac{1 + \epsilon}{1 - \epsilon} |\Delta_S|_1$ , where  $S := \{(i, j) : \Phi_{0ij} \neq 0, i \neq j\}$  and  $S^c := \{(i, j) : \Phi_{0ij} = 0, i \neq j\}$ , holds on  $\mathcal{T}(B_0)$ , and

$$|\Delta_{B_0}|_{1,\text{off}} \leq \frac{1 + \epsilon}{1 - \epsilon} 9 \lambda_A \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1 \sqrt{|B_0^{-1}|_{0,\text{off}}} / \varphi_{\min}^2(\rho(B_0)).$$

Variants of Theorem 4.3 were shown in Rothman et al. (2008) in the context of Gaussian graphical models. We give a complete proof of Theorem 4.3 and Corollary 4.4 in Section D. Lemma 4.5 justifies the choices of the penalty parameters  $\lambda_{A_0}$  and  $\lambda_{B_0}$  as defined in Theorem 3.1,

**Lemma 4.5.** *Let  $\alpha_n, \beta_n < 1/3$  be as defined in Theorem 4.1. Let*

$$\delta_{n,f} = \frac{2\beta_n}{1 - \beta_n} =: C_3 \frac{\log^d(m \vee f)}{\sqrt{nf}} \quad \text{and} \quad \delta_{n,m} = \frac{2\alpha_n}{1 - \alpha_n} =: C_4 \frac{\log^d(m \vee f)}{\sqrt{nm}}.$$

*Then, event  $\mathcal{T}(A_0) \cap \mathcal{T}(B_0)$  holds on  $\mathcal{X}_0$  for the sample correlation matrices as defined in (6a) and (6b) respectively.*

*By Theorem 4.1, we have  $\mathbb{P}(\mathcal{T}(A_0) \cap \mathcal{T}(B_0)) \geq 1 - \frac{3}{(m \vee f)^2}$ .*

**Remark 4.6.** *In Theorem 3.1, we obtain the penalized estimators  $\hat{B}$  and  $\hat{A}$  as defined in (2b) and (2a) with  $\hat{\Gamma}(B_0)$  and  $\hat{\Gamma}(A_0)$  constructed as in (6b) and (6a) as the input; the penalty parameters are chosen to be:*

$$\lambda_{A_0} \asymp C_A \log^d \max(m, f) / \sqrt{mn} \quad \text{and} \quad \lambda_{B_0} \asymp C_B \log^d \max(m, f) / \sqrt{fn},$$

*which on event  $\mathcal{X}_0$  dominate the maximum entry-wise errors  $|\hat{\Gamma}(B_0) - \rho(B_0)|_{\max}$  and  $|\hat{\Gamma}(A_0) - \rho(A_0)|_{\max}$  respectively. The values of  $C_A := \frac{\sqrt{m}\|A_0\|_F}{\text{tr}(A_0)} = \frac{\sqrt{\text{tr}(A_0 A_0)}}{\text{tr}(A_0)}$  and  $C_B := \frac{\sqrt{f}\|B_0\|_F}{\text{tr}(B_0)} = \frac{\sqrt{\text{tr}(B_0 B_0)}}{\text{tr}(B_0)}$  reflect how eigenvalues of each component covariance matrix vary across its entire spectrum. The notation  $\lambda_{A_0}$  and  $\lambda_{B_0}$  thus reflect their dependencies on the eigenspectrum of  $A_0$  and  $B_0$ . Under the bounded spectrum assumptions in (A2),  $C_A, C_B$  are taken to be constants. These choices satisfy the conditions in Theorem 4.3 in view of Lemma 4.5.*

## 5 Variations on a theme

It is curious whether or not one can improve upon the Gemini correlation estimators, which will lead to better convergence rates for estimating  $\rho(A_0)$  and  $\rho(B_0)$  with (2a) and (2b). To answer this question, we introduce a natural variation of the Gemini estimators as given by the Non-iterative Penalized Flip-Flop algorithm described below. To make our discussion concrete, suppose we aim to estimate  $A_* = (a_{*,ij}) = mA_0/\text{tr}(A_0)$  and  $B_* = (b_{*,ij}) = B_0\text{tr}(A_0)/m$  instead of  $A_0$  and  $B_0$ . Note that  $A_*$  has been normalized to have  $\text{tr}(A_*) = m$  for identifiability. Clearly  $A_* \otimes B_* = A_0 \otimes B_0$ . Denote the  $k, \ell^{\text{th}}$  block of size  $f \times f$  in  $\hat{S}_n$  by  $\hat{S}_n^{k\ell}$  and that of size  $m \times m$  in  $\tilde{S}_n$  by  $\tilde{S}_n^{k\ell}$ . We now describe this procedure.

### Non-iterative Penalized Flip-Flop algorithm (NiPFF)

1. Assume  $f \leq m$ . Initialize  $A_{\text{init}} = I$ . Compute  $\hat{\Gamma}(B_0)$  based on (6b) as before, and compute  $\hat{B}$  using GLasso (2b) with the penalty parameter  $\lambda_{A_0}$  to be chosen (cf. Lemma 6.1).

Let  $B_1 = \hat{W}_2 \hat{B} \hat{W}_2 / m$ .

2. Now compute the sample covariance  $\tilde{A}(B_1)$  using (9):

$$\tilde{A}(B_1) = \frac{1}{f} \sum_{k=1}^f \sum_{\ell=1}^f \tilde{S}_n^{k\ell} B_{1,\ell k}^{-1}. \quad (25)$$

Compute the sample correlation matrix  $\hat{\Gamma}(A_0)$  with

$$\hat{\Gamma}(A_0) = \tilde{W}_1^{-1} \tilde{A}(B_1) \tilde{W}_1^{-1} \quad \text{where} \quad \tilde{W}_1 = \text{diag}(\tilde{A}(B_1))^{1/2}. \quad (26)$$



Obtain an estimate  $\hat{A}(B_1)$  using GLasso (2a) with  $\hat{\Gamma}(A_0)$  in (26) as its input, where  $\lambda_B = \lambda_{B_1}$  is to be specified (cf. Remark 6.3 and Corollary 6.4).

Let  $A_1 = \hat{A}_* = \tilde{W}_1 \hat{A}(B_1) \tilde{W}_1$ .

3. Compute sample covariance matrix  $\tilde{B}(A_1)$  using (9):

$$\tilde{B}(A_1) = \frac{1}{m} \sum_{k=1}^m \sum_{j=1}^m \hat{S}_n^{kj} A_{1,jk}^{-1} \quad (27)$$

Compute the sample correlation matrix  $\hat{\Gamma}(B_0)$  with

$$\hat{\Gamma}(B_0) = \tilde{W}_2^{-1} \tilde{B}(A_1) \tilde{W}_2^{-1} \quad \text{where} \quad \tilde{W}_2 := \text{diag}(\tilde{B}(A_1))^{1/2} \quad (28)$$

Obtain an estimate  $\hat{B}(A_1)$  using (2b), with  $\hat{\Gamma}(B_0)$  in (28) as its input, where  $\lambda_A = \lambda_{A_1}$  is to be specified (cf. Theorem 6.6 and Remark 6.7).

Let  $\hat{B}_* = \tilde{W}_2 \hat{B}(A_1) \tilde{W}_2$ .

## 6 Analysis for the penalized Flip-Flop algorithm

In analyzing the Flip-Flop algorithm, we make the following additional assumption.

(A3) The inverse correlation matrices have bounded  $|\rho(A_0)^{-1}|_1$  and  $|\rho(B_0)^{-1}|_1$ :

$$|\rho(A_0)^{-1}|_1 \asymp m \quad \text{and} \quad |\rho(B_0)^{-1}|_1 \asymp f.$$

We use the following notation throughout this section. Let  $c = \frac{\log n}{\log \log(m \vee f)}$  and  $d = 1 - (c/2 \wedge 1/2)$ . Let

$$\lambda_{f,n} = 20 \frac{\log^d \max(m, f)}{\sqrt{fn}} \quad \text{and} \quad \lambda_{m,n} = 20 \frac{\log^d \max(m, f)}{\sqrt{mn}}, \quad (29)$$

and for  $n > 4(C_1 \kappa^2 + C_2 \kappa) \log \max(m, f)$ , where  $\kappa > 0$ ,  $d = 1/2$  and the constant 20 can be replaced with  $2\sqrt{C_1 + C_2/\kappa}$ , where  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ , and  $C_2 = \sqrt{8e} \approx 7.6885$ .

First we bound the entry-wise errors for the sample covariance and correlation matrices as defined in Step 2 in Lemma 6.1 and Theorem 6.2.

**Lemma 6.1.** *Suppose that (A1), (A2) and (A3) hold. Let  $\hat{B}$  and  $B_1$  be obtained as in Step 1, where we choose for  $0 < \varepsilon < 2/3$ ,*

$$\lambda_{A_0} = \frac{2\alpha}{\varepsilon(1-\alpha)} \geq \frac{3\alpha}{1-\alpha} \quad \text{for} \quad \alpha = C_A \lambda_{m,n} \quad \text{where} \quad C_A = \|A_0\|_F \sqrt{m}/\text{tr}(A_0).$$

Then on event  $\mathcal{A}_1$ , for  $\tilde{A}(B_1)$  as defined in (25)

$$\left| \left( \tilde{A}(B_1) - A_* \right)_{ij} \right| \leq \sqrt{a_{*,ii} a_{*,jj}} \lambda_{f,n} (1 + o(1)) + |a_{*,ij}| \tilde{\mu}, \quad (30)$$

$$\text{where} \quad \tilde{\mu} = \lambda_{A_0} \left| \hat{B}^{-1} \right|_{1,\text{off}} / f + \frac{\alpha}{1-\alpha} \left| \hat{B}^{-1} \right|_1 / f \leq \mu \quad (31)$$

$$\text{for} \quad \mu = \lambda_{A_0} |\rho(B_0)^{-1}|_{1,\text{off}} / f + \frac{\alpha}{(1-\alpha)} |\rho(B_0)^{-1}|_1 / f + o(\lambda_{A_0}). \quad (32)$$

Moreover, we have for some constant  $d \leq 8$ ,  $\mathbb{P}(\mathcal{A}_1) \geq 1 - \frac{d}{(m \vee f)^2}$ .

**Theorem 6.2.** Suppose all conditions in Lemma 6.1 hold. Let  $\widehat{\Gamma}(A_0)$  be as defined in (26). Then on event  $\mathcal{A}_1$ , for  $\tilde{\eta} := \lambda_{f,n}(1 + o(1)) + \tilde{\mu}$ ,  $\forall i \neq j$

$$\left| \widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \right| \leq \frac{\lambda_{f,n}(1 + o(1))}{1 - \tilde{\eta}} (1 + |\rho_{ij}(A_0)|) + |\rho_{ij}(A_0)| \frac{2\tilde{\mu}}{1 - \tilde{\eta}}, \quad (33)$$

where  $\tilde{\mu}$  as defined in (31). Then on event  $\mathcal{A}_1$ , for  $\mu$  as in (32)

$$\left| \widehat{\Gamma}(A_0) - \rho(A_0) \right|_{\max} \leq \frac{2\eta}{1 - \eta} \quad \text{where} \quad \eta = \lambda_{f,n}(1 + o(1)) + \mu. \quad (34)$$

**Remark 6.3.** On event  $\mathcal{A}_1$ , the random quantities  $\tilde{\mu}$  and  $\tilde{\eta}$  are upper bounded by  $\mu$  (32) and  $\eta$  (34) respectively, which can be rewritten as follows.

$$\begin{aligned} \text{Define } C_f &:= |\rho(B_0)^{-1}|_1 / f + \frac{2}{\varepsilon} |\rho(B_0)^{-1}|_{1,\text{off}} / f \quad \text{so that} \\ \mu &= \frac{\alpha}{(1 - \alpha)} (C_f + o(1)) \quad \text{and} \quad \eta = (\lambda_{f,n} + \frac{\alpha}{(1 - \alpha)} C_f)(1 + o(1)), \end{aligned}$$

which suggests that we set the penalty in Step 2 in the following order,

$$\lambda_{B_1} \asymp \frac{2\eta}{(1 - \eta)} \asymp \lambda_{f,n} + \frac{\alpha}{1 - \alpha} C_f \asymp \lambda_{f,n} + \lambda_{m,n}.$$

Clearly  $C_f \asymp 1$  under Assumption (A3). We compare the bound in (33) with that of Theorem 4.1 in Section 6.1. We note that the conclusions of Lemma 6.1 and Theorem 6.2 continue to hold even if  $\varepsilon$  is chosen outside of the interval  $(0, 2/3]$ , so long it is bounded away from 0 and 1.

We now compute the rates of convergence in the operator and the Frobenius norm for estimating  $A_*$  with  $\widehat{A}_*$  in Step 2 in Corollary 6.4. The rates we obtain in Corollary 6.4 correspond to exactly those in Corollary A.3 for the Gemini estimator, with slightly better leading constants, where we replace  $\lambda_{B_1}$  with  $\lambda_{B_0}$ . We compare these two penalty parameters in Section 6.1.

**Corollary 6.4.** Suppose (A1), (A2), and (A3) hold. Assume that  $\eta \leq 1/4$ . Suppose that on event  $\mathcal{A}_1$ , we choose for  $\tilde{\eta}$  as defined in Theorem 6.2,

$$\lambda_{B_1} = 2\tilde{\eta}/(\varepsilon_1(1 - \tilde{\eta})), \quad \text{where } 0 < \varepsilon_1 < 1; \quad (35)$$

Then for  $\widehat{A}_*$  as constructed in Step 2 and some constant  $9 < C < 18$ ,

$$\begin{aligned} \left\| \widehat{A}_* - A_* \right\|_2 &\leq 2C\lambda_{B_1} a_{*,\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1, \\ \left\| \widehat{A}_* - A_* \right\|_F &\leq 2C\lambda_{B_1} a_{*,\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m, \\ \left\| \widehat{A}_*^{-1} - A_*^{-1} \right\|_2 &\leq C\lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1/(a_{*,\min} \varphi_{\min}^2(\rho(A_0))), \\ \text{and } \left\| \widehat{A}_*^{-1} - A_*^{-1} \right\|_F &\leq C\lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m/(a_{*,\min} \varphi_{\min}^2(\rho(A_0))) \end{aligned}$$

where  $a_{*,\max} = \max_i a_{*,ii}$  and  $a_{*,\min} = \min_i a_{*,ii}$ .

Next we bound the entry-wise errors for the sample covariance and correlation matrices as defined in Step 3 in Lemma 6.5 and Theorem 6.6.

**Lemma 6.5.** Suppose that (A1), (A2), and (A3) hold. Let  $\hat{A} = \hat{A}(B_1)$  and  $A_1$  be defined as in Step 2. Suppose we choose  $\lambda_{B_1}$  as in (35). Then on event  $\mathcal{A}_1 \cap \mathcal{E}_2$ , for  $\tilde{B}(A_1)$  as in (27) and  $\eta$  as in (34),

$$\left| \left( \tilde{B}(A_1) - B_* \right)_{ij} \right| \leq \sqrt{b_{*,ii} b_{*,jj}} \lambda_{m,n} (1 + o(1)) + |b_{*,ij}| \tilde{\xi} \quad (36)$$

$$\text{where } \tilde{\xi} = \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}} / m + \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \hat{A}^{-1} \right|_1 / m \leq \xi$$

$$\text{for } \xi = \lambda_{B_1} \left| \rho(A_0)^{-1} \right|_{1,\text{off}} / m + \frac{\eta}{1 - \eta} \left| \rho(A_0)^{-1} \right|_1 / m + o(\lambda_{B_1}). \quad (37)$$

Moreover, we have for some constant  $d \leq 10$ ,  $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{E}_2) \geq 1 - \frac{d}{(m \vee f)^2}$ .

**Theorem 6.6.** Suppose all conditions in Lemma 6.5 hold. Let  $\zeta = \lambda_{m,n} (1 + o(1)) + \xi$ , where  $\xi$  is as defined in (37). Then on event  $\mathcal{A}_1 \cap \mathcal{E}_2$ ,

$$\forall i \neq j \quad \left| \hat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \right| \leq \frac{\lambda_{m,n} (1 + o(1))}{1 - \zeta} + |\rho_{ij}(B_0)| \frac{\zeta + \xi}{1 - \zeta} \quad (38)$$

$$\leq \frac{2\lambda_{m,n} (1 + o(1))}{1 - \zeta} + |\rho_{ij}(B_0)| \frac{2\xi}{1 - \zeta}. \quad (39)$$

## 6.1 Discussion

We first compare the bound in (33) with that of Theorem 4.1, where we obtain with probability at least  $1 - \frac{3}{\max(m, f)^2}$ ,

$$\forall i \neq j, \quad \left| \hat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \right| \leq \frac{C_B \lambda_{f,n}}{1 - \beta} (1 + |\rho_{ij}(A_0)|). \quad (40)$$

for  $\hat{\Gamma}(A_0)$  as defined in (6a),  $\beta = C_B \lambda_{f,n}$ , and  $C_B = \|B_0\|_F \sqrt{f} / \text{tr}(B_0)$ . On the other hand, the influence of  $\lambda_{A_0} \asymp \frac{2\alpha}{1-\alpha}$  on the entry-wise error for estimating  $\rho_{ij}(A_0)$  in (33) is regulated through both  $C_f$ , which is a bounded constant under (A3) (see Remark 6.3), as well as the magnitude of  $\rho_{ij}(A_0)$  itself; to see this, we write (33) as follows:  $\forall i \neq j$ ,

$$\left| \hat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \right| \leq \left( \frac{\lambda_{f,n}}{1 - \eta} (1 + |\rho_{ij}(A_0)|) + \frac{2\alpha}{1 - \alpha} C_f |\rho_{ij}(A_0)| \right) (1 + o(1)).$$

We note that this rate is at the same order as that in (40). However, when  $\lambda_{m,n} \ll \lambda_{f,n}$ , the second term  $\frac{2\alpha}{1-\alpha} C_f |\rho_{ij}(A_0)| \asymp |\rho_{ij}(A_0)| C_f C_A \lambda_{m,n}$  is of smaller order compared to the first term. In this case, the upper bound in (33) is dominated by the first term on the RHS, and one can perhaps obtain a slightly better bound with Theorem 6.2, as the leading term no longer depends on the constant  $C_B$  as displayed in (40). For example, suppose that  $B_0$  is a diagonal matrix so that  $\rho(B_0)^{-1} = I$ , so that  $|\rho(B_0)^{-1}|_1 / f = 1$  and  $C_f = 1$ . The constant  $C_B = \sqrt{f} \|B_0\|_F / \text{tr}(B_0) \geq 1$  can still be large due to the variability of the diagonal entries of  $B_0$ . This will lead to improvements in estimating  $A_*$  as shown in Corollary 6.4, which remains true so long as  $B_0^{-1}$  is sparse in the sense of (A3).

We next compare the bound in (38) with that of Theorem 4.1, where we obtain with probability at least  $1 - \frac{3}{\max(m, f)^2}$ , for  $\hat{\Gamma}(B_0)$  as defined in (6b),

$$\forall i \neq j, \quad \left| \hat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \right| \leq \frac{C_A \lambda_{m,n}}{1 - \alpha} (1 + |\rho_{ij}(B_0)|) \quad (41)$$

where  $\alpha = C_A \lambda_{m,n}$ , and  $C_A = \|A_0\|_F \sqrt{m}/\text{tr}(A_0)$ . Before we proceed, we first define the following parameter

$$C_m = |\rho(A_0)^{-1}|_1/m + \frac{2}{\varepsilon_1} |\rho(A_0)^{-1}|_{1,\text{off}}/m \quad \text{so that}$$

$$\xi \leq \frac{\eta}{1-\eta} (C_m + o(1)) \quad \text{and} \quad \zeta \leq (\lambda_{m,n} + \frac{\eta}{1-\eta} C_m)(1 + o(1))$$

where  $0 < \varepsilon_1 < 1$  is the same as in (35). Hence we have on  $\mathcal{A}_1 \cap \mathcal{E}_2$ ,

$$\begin{aligned} \left| \widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \right| &\leq \frac{\lambda_{m,n}(1 + o(1))}{1 - \zeta} + |\rho_{ij}(B_0)| \frac{\zeta + \xi}{1 - \zeta} \\ &\leq \frac{3}{2} \left( \lambda_{m,n}(1 + |\rho_{ij}(B_0)|) + |\rho_{ij}(B_0)| C_m \frac{2\eta}{1 - \eta} \right) (1 + o(1)) \end{aligned}$$

by (38), where  $2\eta/(1 - \eta) \asymp \lambda_{m,n} + \lambda_{f,n}$  and we assume that  $\zeta < 1/3$ . Clearly the influence of  $\lambda_{B_1} \asymp \frac{2\eta}{1-\eta}$  on the entry-wise error for estimating  $\rho_{ij}(B_0)$  is regulated through the quantity  $C_m$  which is a constant under (A3), as well as the magnitude of  $\rho_{ij}(B_0)$  itself. We note that these rates are at the same order as in Theorem 4.1 when  $m \asymp f$ . Moreover, for pairs of  $(i, j)$  where  $i \neq j$ , such that  $|\rho_{ij}(B_0)|$  is small, one can perhaps obtain a slightly better bound with Theorem 6.6, as the first (leading) term no longer depends on the constant  $C_A = \sqrt{m} \|A_0\|_F / \text{tr}(A_0) \geq 1$  as needed in (41). Otherwise, suppose that  $m \gg f$ . Then the original estimator in (6b) could be much better for pairs of  $(i, j)$  with a large  $|\rho_{ij}(B_0)|$ , as for such pairs, the second term is of larger order than the first term in (39).

**Remark 6.7.** Finally, we mention that for the  $\widehat{B}_*$  which we obtain in Step 3, we achieve the same convergence bounds as in Corollary A.3, except that we replace  $\lambda_{A_0}$  with  $\lambda_{A_1}$ , which will be chosen to dominate the maximum entry-wise error: assuming that  $\zeta < 1/3$  and  $\eta < 1/4$ ,

$$\left| \widehat{\Gamma}(B_0) - \rho(B_0) \right|_{\max} \leq 3\lambda_{m,n} + 4 \max_{i \neq j} |\rho_{ij}(B_0)| C_m (\lambda_{f,n} + C_f C_A \lambda_{m,n} / (1 - \alpha)).$$

In summary, for the following cases, we expect that the estimate  $\widehat{\Gamma}(B_0)$  which we obtain in Step 3 improves upon the initial estimate in Step 1. Let  $\zeta' = \lambda_{m,n}(1 + o(1)) + \max_{i \neq j} |\rho_{ij}(B_0)| \xi$ . Clearly the RHS of (39) is bounded by  $2\zeta'/(1 - \zeta)$ .

1. For all  $i \neq j$ ,  $\rho_{ij}(B_0)$  is bounded in magnitudes; For example, when  $|\rho_{ij}(B_0)| = O(\sqrt{f/m})$ , then  $\zeta' \asymp \lambda_{m,n}$ . In particular,

$$\forall i \neq j, \quad \left| \widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \right| \leq \frac{2\zeta'}{1 - \zeta} \approx \frac{2\lambda_{m,n}(1 + o(1))}{1 - \zeta} \leq 3\lambda_{m,n}$$

if  $B_0$  is a diagonal matrix. Hence the error in estimating  $A_0$  is propagated into the estimate of  $\rho_{ij}(B_0)$  only when  $\rho_{ij}(B_0) \neq 0$ .

2. When  $m$  and  $f$  are close to each other in that the ratio  $m/f \rightarrow \text{const} > 1$ , and simultaneously,  $C_m$ ,  $C_f$ , and  $|\rho_{ij}(B_0)|$  are small for all  $i \neq j$ ; Then  $\zeta' \asymp \lambda_{m,n} + \lambda_{f,n}$  provides a tight bound.

A refined analysis on the GLasso given the estimates in Theorem 6.6 is left as future work.

## 7 Numerical results

We demonstrate the effectiveness of the Gemini method as well as the Non-iterative penalized Flip-Flop method, which we refer to as the FF method, with simulated data. For a penalty parameter  $\lambda \geq 0$ , the GLasso estimator is given by

$$\text{glasso}(\hat{\Gamma}, \lambda) = \underset{\Theta \succ 0}{\operatorname{argmin}} (\operatorname{tr}(\hat{\Gamma}\Theta) - \log |\Theta| + \lambda |\Theta|_{1,\text{off}}),$$

where  $\hat{\Gamma}$  is a sample correlation matrix. We use the R-package `glasso` [Friedman et al. \(2008\)](#) to compute the GLasso solution. For the two estimation methods we have various tuning parameters, namely  $\lambda, \nu$  for the baseline Gemini estimators, and  $\phi, \nu$  for the FF method. In our simulation study, we look at three different models from which  $A$  and  $B$  will be chosen. Let  $\Omega = A^{-1} = (\omega_{ij})$  and  $\Pi = B^{-1} = (\pi_{ij})$ . Let  $E$  denote edges in  $\Omega$ , and  $F$  denote edges in  $\Pi$ . We choose  $A$  from one of these two models:

- AR(1) model. In this model the covariance matrix is of the form  $A = \{\rho^{|i-j|}\}_{i,j}$ . The graph corresponding to  $\Omega$  is a chain.
- Star-Block model. In this model the covariance matrix is block-diagonal with equal-sized blocks whose inverses correspond to star structured graphs, where  $A_{ii} = 1$ , for all  $i$ . We have 20 subgraphs, where in each subgraph, 8 nodes are connected to a central hub node with no other connections. The rest of the nodes in the graph are singletons. Covariance matrix for each block  $S$  in  $A$  is generated as in [Ravikumar et al. \(2008\)](#):  $S_{ij} = \rho = 0.5$  if  $(i, j) \in E$  and  $S_{ij} = \rho^2$  otherwise.

For  $\Pi$ , we use the random concentration matrix model in [Zhou et al. \(2008\)](#). The graph is generated according to a type of Erdős-Rényi random graph model. Initially we set  $\Pi = 0.25I_{f \times f}$ , where  $f = 80$ . Then, we randomly select  $d$  edges and update  $\Pi$  as follows: for each new edge  $(i, j)$ , a weight  $w > 0$  is chosen uniformly at random from  $[w_{\min}, w_{\max}]$  where  $w_{\max} > w_{\min} > 0$ ; we subtract  $w$  from  $\pi_{ij}$  and  $\pi_{ji}$ , and increase  $\pi_{ii}$  and  $\pi_{jj}$  by  $w$ . This keeps  $\Pi$  positive definite. For both models of  $A$ , we have  $A_* = A \frac{m}{\operatorname{tr}(A)} = A = \rho(A)$ . Let  $\Omega_* = \frac{\operatorname{tr}(A)}{m} \Omega$  and  $\Pi_* = \frac{m}{\operatorname{tr}(A)} \Pi$ . Thus we have  $\Omega_* = \Omega$  and  $\Pi_* = \Pi$  for all combinations of  $A$  and  $B$  in this section.

### 7.1 Regularization Paths and Cross-validation

We illustrate the behaviors of the Gemini estimators for each model combination of  $A, B$  with  $m = 400$  and  $f = 80$  over the full regularization paths. To evaluate consistency, we use relative errors in the operator and the Frobenius norm. For model selection consistency, we use false positive and false negative rates defined in Table 1. For each pair of covariance matrices, we do the following. First, we generate  $A$  and  $B$ , where  $A$  is  $m \times m$  and  $B$  is  $f \times f$ . Let  $A^{1/2}$  and  $B^{1/2}$  be the unique square root of matrix  $A$  and  $B$  respectively. Let  $T$  and  $T'$  be a set of values in  $(0, 0.5]$ . Now, repeat the following steps 100 times:

1. Sample random matrices  $X^{(1)}, \dots, X^{(n)} i.i.d. \sim \mathcal{N}_{f,m}(0, A \otimes B)$ :

$$X^{(t)} = B^{1/2} Z(t) A^{1/2}, \text{ where } Z_{ij}(t) \sim N(0, 1) \quad \forall i, j, \forall t = 1, \dots, n.$$

Compute the sample column correlation  $\widehat{\operatorname{corr}}_{\text{col}}$  as in (6a) and row correlation  $\widehat{\operatorname{corr}}_{\text{row}}$  as in (6b).

2. For each  $\lambda \in T$  and  $\nu \in T'$ :

- (a) Obtain the estimated inverse correlation matrices  $\hat{A}^{-1}$ , and  $\hat{B}^{-1}$  with  $\text{glasso}(\widehat{\text{corr}}_{\text{col}}, \lambda)$  and  $\text{glasso}(\widehat{\text{corr}}_{\text{row}}, \nu)$  respectively. Let  $\hat{\Omega}(\lambda) := \hat{A}_*^{-1}$  and  $\hat{\Pi}(\nu) := \hat{B}_*^{-1}$ , where  $\hat{A}_*$  and  $\hat{B}_*$  are as defined in (17).
- (b) Let  $\hat{E}(\lambda)$  denote the set of edges in the estimated  $\hat{\Omega}(\lambda)$ . Now compute  $\text{FNR}(\lambda)$  and  $\text{FPR}(\lambda)$  as defined in Table 1. To obtain  $\text{FNR}(\nu)$  and  $\text{FPR}(\nu)$ , we need to replace  $\hat{E}(\lambda)$  with  $\hat{F}(\nu)$ , which denotes the set of edges in  $\hat{\Pi}(\nu)$ ,  $E$  with  $F$ , and  $m$  with  $f$ . Compute the relative errors  $\|\hat{\Omega}(\lambda) - \Omega\| / \|\Omega\|$  and  $\|\hat{\Pi}(\nu) - \Pi\| / \|\Pi\|$ , where  $\|\cdot\|$  denotes either the operator or the Frobenius norm.

Table 1: Metrics for evaluating  $\hat{E}(\lambda)$

Metric	Definition
False Positives (FPs)	# of incorrectly selected edges in $\hat{E}(\lambda)$ : $ \hat{E}(\lambda) \setminus E $
False Negatives (FNs)	# of edges in $E$ that are not selected in $\hat{E}(\lambda)$ : $ E \setminus \hat{E}(\lambda) $
True positives (TPs)	# of correctly selected edges: $ \hat{E}(\lambda) \cap E $
True Negatives (TNs)	# of zeros in $\hat{E}(\lambda)$ that are also zero in $E$
False Positive Rate (FPR)	$\text{FPR} = FP / (FP + TN) = FP / (\binom{m}{2} -  E )$
False Negative Rate (FNR)	$\text{FNR} = FN / (TP + FN) = FN /  E $

After 100 trials, we plot each of the following as  $\lambda$  changes over a range of values in  $T$ :  $(\overline{\text{FNR}} + \overline{\text{FPR}})(\lambda)$  for  $\hat{E}(\lambda)$ , where  $\overline{\text{FNR}}$ ,  $\overline{\text{FPR}}$  are averaged over the 100 trials for each metric, and the average relative errors in the operator and the Frobenius norm. Similarly, we plot these as  $\nu$  changes over a range of values in  $T'$ . Figure 2 shows how these three metrics change as the  $\ell_1$  regularization parameters  $\lambda$  and  $\nu$  increase over full paths where covariance  $A$  comes from either AR(1) or the Star-Block model, and  $\Pi$  comes from the random graph model. These plots show that the Gemini method is able to select the correct structures as well as achieving low relative errors in the operator and the Frobenius norm when  $\lambda$  and  $\nu$  are chosen from a suitable range. In addition, as  $n$  increases, we see performance gains over almost the entire paths for all three metrics as expected. Other model combinations of  $A, B$  which are not shown here confirm similar findings.

In Figure 2, we also illustrate choosing the penalty parameters  $\lambda$  and  $\nu$  by 10-fold cross-validation. To do so, we run the following for 10 trials. In each trial, we partition the rows of each  $X^{(t)}$ ,  $t = 1, \dots, n$  into 10 folds. For each fold, the validation set consists of the subset of rows of  $X^{(1)}, \dots, X^{(n)}$  sharing the same indices and its complement set serves as the training data. Denote by  $\widehat{\text{corr}}_T$  and  $\widehat{\text{corr}}_V$  the column-wise sample correlations based upon the training and the validation data, which are computed in the same manner as in (6a). We define  $\text{score}_A(\lambda) = \text{tr}(\hat{\Theta}_\lambda \widehat{\text{corr}}_V) - \log |\hat{\Theta}_\lambda|$ , where  $\hat{\Theta}_\lambda = \text{glasso}(\widehat{\text{corr}}_T, \lambda)$ . The final score for a particular  $\lambda$  is the average over 10 trials (with 10 folds in each trial) and the one with the lowest score is chosen to be  $\lambda_{CV}$ . Similarly, we use column partitions to obtain  $\nu_{CV}$ .

## 7.2 ROC comparisons

In this section, we compare the performances of the two methods, namely, the baseline Gemini and its three-step FF variant over the full paths by examining their ROC curves. Each curve is an average over 50 trials. We fix  $f = 80$ ,  $m = 400$ ,  $n = 1$ . To simplify our notation, we summarize the penalty parameters which we use for indexing the ROC curves as follows:

$$\lambda = \lambda_{B_0} \quad \nu = \lambda_{A_0} \quad \phi = \lambda_{B_1} \quad v = \lambda_{A_1}$$

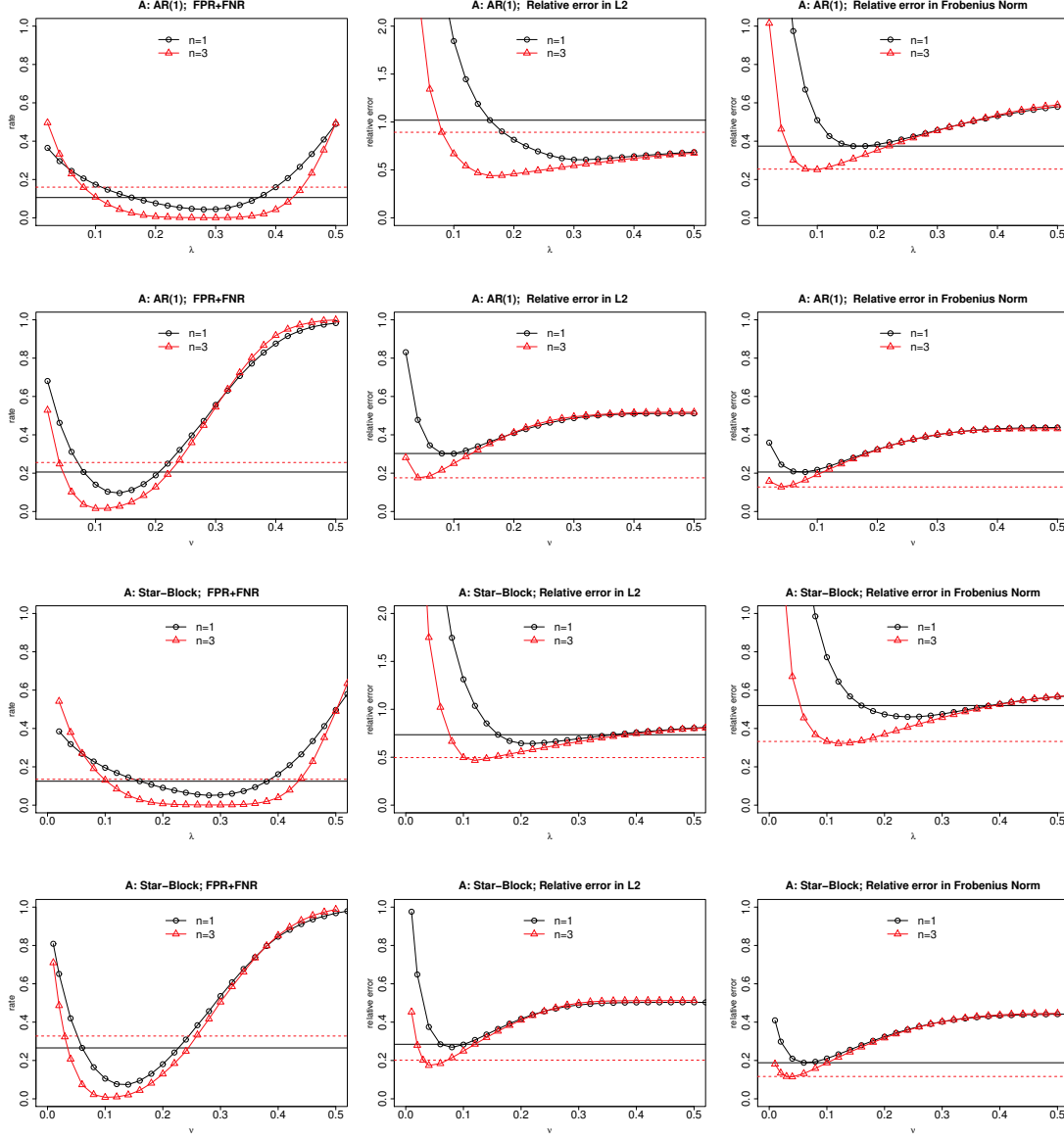


Figure 2:  $m = 400$ ,  $f = 80$ ,  $n = 1$ .  $B^{-1}$  follows random graph model with  $d = 80$  and  $w \in [0.1, 0.3]$  throughout these plots. In the top two panels, covariance  $A$  follows the AR(1) model with  $\rho = 0.5$ ; for the bottom two panels,  $A$  follows the Star-Block model. The top and the third panel are for  $\hat{\Omega}(\lambda)$ ; the second and the bottom panel are for  $\hat{\Pi}(\nu)$ . As penalization parameter  $\lambda$  or  $\nu$  increases, FPs decrease and FNIs increase; therefore, we see the bowl-shaped curves in all plots on the left column. The relative errors also first decrease and then increase before they level off. This happens because at first, decreased FPs clearly help to reduce the estimation errors; however, as penalization further increases, the estimated graphs are missing more and more edges until the inverse covariance estimates have only diagonal entries in the end. Solid and dashed horizontal lines show the performances of Gemini for cross-validated tuning parameters: in the top two panels,  $\lambda_{CV} = 0.16$  and  $\nu_{CV} = 0.08$  for  $n = 1$ , and  $\lambda'_{CV} = 0.08$  and  $\nu'_{CV} = 0.04$  for  $n = 3$  respectively. For the bottom two panels,  $\lambda_{CV} = 0.16$  and  $\nu_{CV} = 0.10$  for  $n = 1$ , and  $\lambda'_{CV} = 0.06$  and  $\nu'_{CV} = 0.03$  for  $n = 3$  respectively. The cross-validated tuning parameters tend to stay close to the point of  $\lambda$  or  $\nu$  which minimizes the relative error in the Frobenius norm.



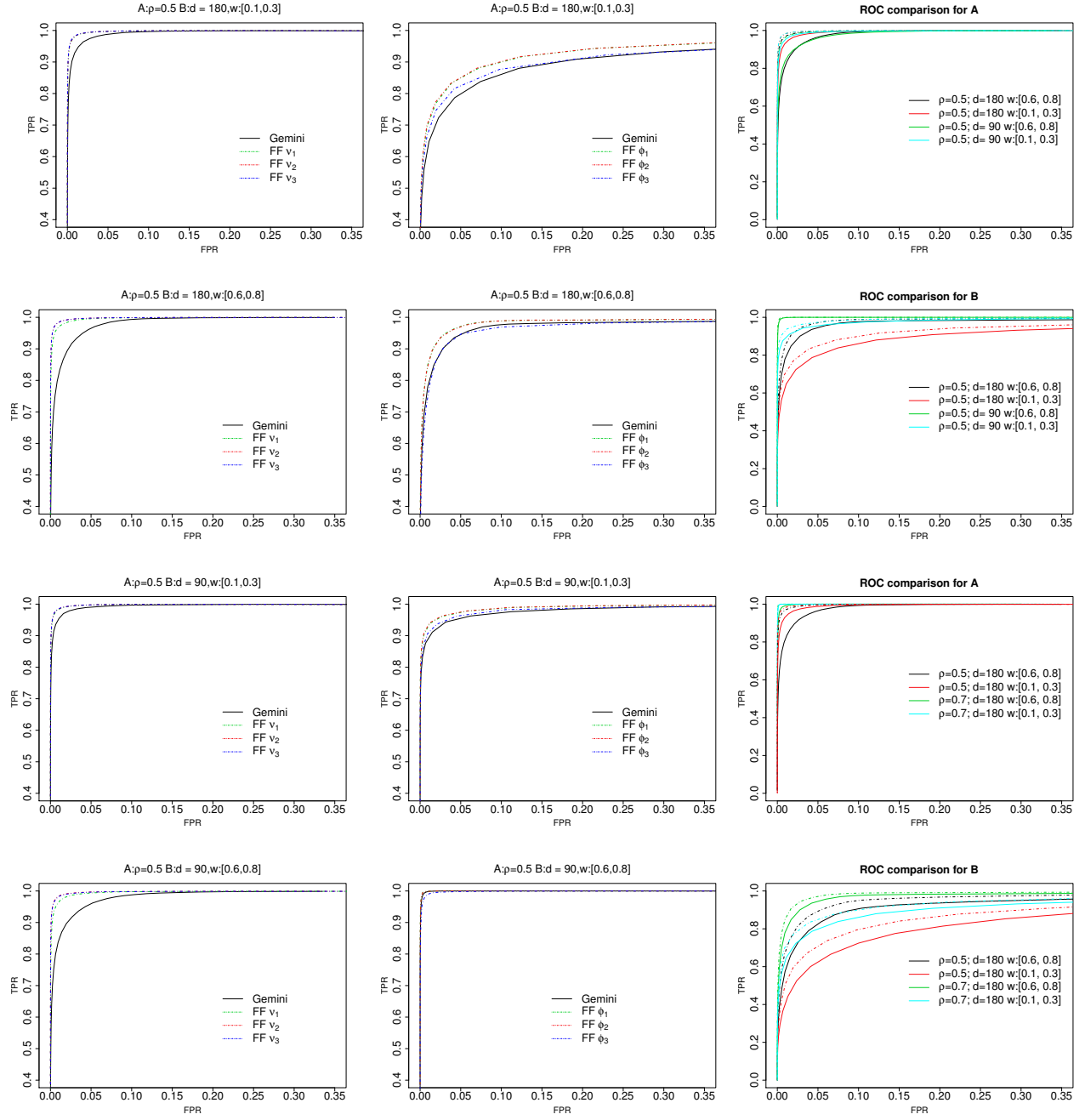


Figure 3:  $m = 400, f = 80, n = 1$ . Solid lines are for Gemini. Plots in the left column are for  $A$  and the middle column are for  $B$ . The three dotted lines in each plot on the left column correspond to the three optimization criteria  $\nu_1, \nu_2, \nu_3$  as specified in Step 3. For the middle column, they correspond to  $(i_1, \phi_1), (i_2, \phi_2), (i_3, \phi_3)$ , as specified in Step 4. In the right column: in top two plots, we choose  $A$  from AR(1) model with  $\rho = 0.5$  while changing the settings of  $B^{-1}$  as in Table 2; In bottom two plots, we choose  $A$  from AR(1) model with  $\rho = 0.5$  or  $0.7$  while changing the settings of  $B^{-1}$  with  $d = 180$ . Dotted lines in plots for  $A$  on the right column are chosen according to the optimization criterion  $\nu_1$ , and in plots for  $B$ , they are chosen according to the criterion  $\phi_1$ .

To illustrate the overall performances of the baseline Gemini method for estimating the graphs of  $\Omega$  and  $\Pi$ , we use pairs of metrics  $(\overline{\text{FPR}}(\lambda), 1 - \overline{\text{FNR}}(\lambda))$  and  $(\overline{\text{FPR}}(\nu), 1 - \overline{\text{FNR}}(\nu))$  respectively, which we obtain as the average over 50 trials of Step 1 and 2 as described in Section 7.1. To plot the ROC curves for the FF method, we start with estimating  $\Pi$  with the Gemini estimator. Due to computational complexity, we specify the input parameters of the subsequent steps sequentially. These choices are not feasible in practical settings. We run through this idealized example for the sake of comparing with the baseline Gemini estimators. Repeat the following 50 times: Let  $T := \{0.02, 0.04, \dots, 0.72\}$ .

1-2 Run Step 1, 2 as in Section 7.1, while only computing the metrics for  $\widehat{\Pi}(\nu)$ , where  $\nu \in T$ .

3 To execute the second step of the FF algorithm, we use the following three outputs from Step 2 of the current procedure to act as  $B_1$  to compute  $\tilde{A}(B_1)$ . We choose the output  $B_1$  such that its corresponding  $\nu$  is chosen to be  $\nu_1 = \arg \min_{\nu \in T} (\text{FNR} + \text{FPR})(\nu)$ ,  $\nu_2 = \arg \min_{\nu \in T} \|\widehat{\Pi}(\nu) - \Pi\|_2 / \|\Pi\|_2$ , and  $\nu_3 = \arg \min_{\nu \in T} \|\widehat{\Pi}(\nu) - \Pi\|_F / \|\Pi\|_F$ . Denote these by  $B_1^1, B_1^2$  and  $B_1^3$ . We now run the second step of the FF method for each  $B_1^i$ , where  $i = 1, 2, 3$ , with penalty parameter  $\phi \in T$  changing over the full path while obtaining the inverse estimators  $\widehat{\Omega}^i(\phi)$  for  $\Omega$  and computing  $\text{FNR}^i(\phi)$ , and  $\text{FPR}^i(\phi)$  for each estimated edge set. These contribute to 3 ROC curves for estimating the edges in  $E$ .

4 To execute the last step of the FF method, we use the following three outputs from Step 3 as  $A_1$  to compute  $\tilde{B}(A_1)$ . We choose the output  $A_1$  such that its corresponding  $(i, \phi)$  is chosen to be optimal with respect to one of the following metrics:  $(i_1, \phi_1) = \arg \min_{\phi \in T, i=1,2,3} (\text{FNR}^i + \text{FPR}^i)(\phi)$ ,  $(i_2, \phi_2) = \arg \min_{\phi \in T, i=1,2,3} \|\widehat{\Omega}^i(\phi) - \Omega\|_2 / \|\Omega\|_2$ , and  $(i_3, \phi_3) = \arg \min_{\phi \in T, i=1,2,3} \|\widehat{\Omega}^i(\phi) - \Omega\|_F / \|\Omega\|_F$ . The choices then become  $(\nu_{i_j}, \phi_j), j = 1, 2, 3$ , which we simply denote by  $\phi_1, \phi_2, \phi_3$ . Thus, there are again three choices for  $A_1$ . We now run the third step of the FF method for each  $\tilde{B}(A_1)$  with  $v \in T$  changing over the full path, while computing  $\text{FNR}^j(v)$  and  $\text{FPR}^j(v)$ , where  $j = 1, \dots, 3$ , for each estimated edge set. These contribute to 3 ROC curves for estimating the edges of  $F$ .

The ROC curves are plotted in Figure 3 using pairs of metrics  $(\overline{\text{FPR}}^i(\phi), 1 - \overline{\text{FNR}}^i(\phi))$  and  $(\overline{\text{FPR}}^j(v), 1 - \overline{\text{FNR}}^j(v))$ ,  $i, j = 1, 2, 3$ , which are averaged over 50 trials. Throughout the plots on the left column of Figure 3, we see clear performance gains of the FF method over the baseline Gemini on estimating  $\Omega = A^{-1}$ , when the initial penalty  $\nu$  is chosen properly. For  $\Pi = B^{-1}$  in the middle column, we do not always see improvements when  $w$  is drawn from  $[0.6, 0.8]$ . We do see some improvements in case  $w$  is drawn from  $[0.1, 0.3]$  and when the total correlation  $\rho_B^2$  (see definition below) is small. Overall, the performance gains for  $\Pi$  are not as substantial as those for  $\Omega$ . These observations are consistent with our theory and discussion in Section 6.1.

### 7.3 Summary

We use the following metrics to compare matrix  $B$  and  $A$  across different models or parameters:

1. Total correlation:  $\rho_A^2 = \sum_{i < j} \rho_{ij}^2(A) / \binom{m}{2}$  and  $\rho_B^2 = \sum_{i < j} \rho_{ij}^2(B) / \binom{f}{2}$ .
2.  $\|B\|_F / \text{tr}(B)$  and  $\|A\|_F / \text{tr}(A)$ : these affect the entry-wise error bound in sample correlation estimates for  $\rho_{ij}(A)$  and  $\rho_{ij}(B)$ , for all  $i \neq j$ , for the baseline Gemini estimators.
3. The pairs of  $\ell_1$ -metrics  $(|\rho(B)^{-1}|_{1,\text{off}}, |\rho(B)^{-1}|_1)$  and  $(|\rho(A)^{-1}|_{1,\text{off}}, |\rho(A)^{-1}|_1)$ .

The total correlation metric comes from [Efron \(2009\)](#). We use it to characterize the average squared magnitudes for correlation coefficients of  $\rho(A)$  or  $\rho(B)$ . They are clearly relevant for the FF method as the entry-wise error bound for estimating  $\rho_{ij}(A)$  and  $\rho_{ij}(B)$ , for all  $i \neq j$ , depends on the magnitude of the entry itself (cf. Theorem 6.2 and 6.6). We summarize the metrics for  $B$  in table 2.

Table 2: Metrics for comparing the ROC curves

Metric	d=90, w:[0.1,0.3]	d=180, w:[0.1,0.3]	d=90, w:[0.6,0.8]	d=180, w:[0.6,0.8]
$\rho_B^2$	0.053	0.06	0.094	0.12
$\ B\ _F / \text{tr}(B)$	0.128	0.13	0.155	0.16
$\ell_1$ -metrics	(55, 152)	(71, 166)	(99, 225)	(102, 216)

We summarize our findings across the ROC curves in the right column in Figure 3. First we focus on the case when  $A$  is fixed and  $B$  is changing. When  $\Pi$  follows the random graph model, we observe that for both the baseline Gemini estimators and their FF variants, the performances in terms of estimating edges for  $\Omega$  are better when the weights for  $\Pi$  are chosen from  $[0.1, 0.3]$  for both  $d = 90$  and  $d = 180$ . Here the sparsity for  $\Pi$  is not the decisive factor. This is consistent with our theory, in view of Table 2, that  $\|B\|_F / \text{tr}(B)$  affects the entry-wise error bound for the baseline Gemini correlation estimate  $\hat{\Gamma}(A)$  as shown in Corollary 2.2, and the pair of metrics  $(|\rho(B)^{-1}|_{1,\text{off}}, |\rho(B)^{-1}|_1)$  affect that for the FF correspondent in (26) as shown in Theorem 6.2. The performances in terms of edge recovery for  $\Pi$  take a different order. The sparse random graphs with  $d = 90$  see better performances than those with  $d = 180$  for both the Gemini and the FF methods. For graphs with the same sparsity, the one with the larger weight performs better. This is consistent with our theory in Section D.2.

Next we choose two covariance matrices for both  $A$  and  $B$ : for  $B$ , we choose the two cases with different edge weights with  $d = 180$ ; and for  $A$ , we set the parameter  $\rho$  to 0.5 or 0.7. The metrics for the two choices of  $A$  are: for  $\rho = 0.5$ , we have  $\rho_A^2 = 0.04$ ,  $\|A\|_F / \text{tr}(A) = 0.065$ , and  $\ell_1$ -metrics = (532, 1198). The corresponding numbers for  $\rho = 0.7$  are: 0.07, 0.085, and (1095, 2262) respectively.

First we note that the two cases of  $B$  show the same trend when  $A$  is fixed. In the right bottom two plots in Figure 3, for the graphs of  $\Omega$ , we find it easier to estimate when their covariance matrices come with parameter  $\rho = 0.7$ , which results in larger  $\ell_1$  metrics, and hence larger weights on the inverse chain graph; for the graphs of  $\Pi$ , we observe relatively larger performance gains when  $\rho = 0.5$  for  $A$ , with the most significant occurring when  $w \in [0.1, 0.3]$  for  $\Pi$ , where both  $\rho(A)^{-1}$  and  $\rho(B)^{-1}$  have smaller  $\ell_1$  metrics and the total correlation  $\rho_B^2 = 0.06$  is also small. The least improvement we see occurs in case all three metrics are large:  $\rho = 0.7$ ,  $w \in [0.6, 0.8]$ , and  $\rho_B^2 = 0.12$ . These findings are consistent with results in Theorem 6.2 and 6.6, where we explicitly show the influence of the pairs of  $\ell_1$ -metrics on the error bounds for the FF sample correlation estimates.

## 8 Conclusion

In this paper, we presented two methods for estimating the graphs in a matrix variate normal model. The baseline Gemini method is rather simple and provides the same rates of convergence as the Non-iterative Penalized Flip-Flop method in the operator and the Frobenius norm. Under sparsity conditions as detailed in (A1) and (A2), the NiPFF method improves upon the baseline algorithm in estimating  $A_0^{-1}$ , which is

assumed to be the one with the larger dimension, so long as  $\rho(B_0)^{-1}$  satisfies a certain additional sparsity condition, namely, its  $\ell_1$  metrics are bounded in the order of its dimensionality.

It remains an open question whether the iterative methods will help achieve better convergence rates in the one-sample or finite-sample instances. We leave the answer to this question in future work; in the current work, we illustrate that under suitable conditions, our three-step approximation could indeed improve upon the baseline algorithm. However, we show in both theoretical analysis and simulation results that the performance gains for estimating  $B_0^{-1}$  using the NiPFF method at the third step are rather limited; hence we do not advocate iterating beyond the first three steps.

Although our primary interests are in estimating correlations and partial correlations among and between both rows and columns when  $X$  follows a matrix variate normal distribution, our methods clearly can be extended to the general cases when the data matrix  $X$  follows other type of matrix-variate distributions. Theoretical analysis on such generalized models will be left as future work. Finally, Hoff (2011) showed that the class of separable covariance models for random arrays of arbitrary dimension can be generated with a type of multilinear transformation of an array of independent, standard normal entries. It will be interesting to apply our methods to the array normal setting.

## Acknowledgement

The author is grateful for the helpful discussions with Xuming He, John Lafferty, Mark Rudelson, Kerby Shedden, and Stanislaw Szarek.

## References

- ALLEN, G. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics* **4** 764–790.
- BANERJEE, O., GHAOUI, L. E. and D’ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516.
- BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57** 33–45.
- BICKEL, P. and LEVINA, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.
- BONILLA, E., CHAI, K. and WILLIAMS, C. (2008). Multi-task gaussian process prediction. In *In Advances in Neural Information Processing Systems 20 (NIPS 2010)*.
- CAI, T., LIU, W. and LUO, X. (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.
- CAI, T., ZHANG, C.-H. and ZHOU, H. (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics* **38** 2118–2144.

- DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika* **68** 265–274.
- DUTILLEUL, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **64** 105–123.
- EFRON, B. (2009). Are a set of microarrays independent of each other? *Ann. App. Statist.* **3** 922–942.
- EL KAROUI, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics* **36** 2717–2756.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.
- FURRER, R. and BENGTTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* **98** 227–255.
- GUPTA, A. and VARGA, T. (1992). Characterization of matrix variate normal distributions. *Journal of Multivariate Analysis* **41** 80–88.
- HOFF, P. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis* **6** 179–196.
- HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.
- LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* **37** 4254–4278.
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and processes*. Springer.
- LU, N. and ZIMMERMAN, D. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.
- PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104** 735–746.
- RAUHUT, H., SCHNASS, K. and VANDERGHEYNST, P. (2008). Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory* **54** 2210–2219.
- RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. In *Advances in Neural Information Processing Systems*. MIT Press.
- ROTHMAN, A., BICKEL, P., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

- TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal Optimization Theory and Applications* **109** 475 – 494.
- TSILIGKARIDIS, T., HERO, A. and ZHOU, S. (2012). Convergence properties of kronecker graphical lasso algorithms. In submission. Available at <http://arxiv.org/1204.0585v1.pdf>.
- VERSHYNIN, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press. Compressed Sensing, Theory and Applications.
- WEICHSEL, P. (1962). The kronecker product of graphs. *Proc. Amer. Math. Soc.* **13** 47–52.
- WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing* **56** 478 – 491.
- YIN, J. and LI, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis* **107** 119–140.
- YU, K., LAFFERTY, J., ZHU, S. and GONG, Y. (2009). Large-scale collaborative prediction using a non-parametric random effects model. *Proceedings of the 26th International Conference on Machine Learning*.
- YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11** 2261–2286.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.
- ZHANG, Y. and SCHNEIDER, J. (2010). Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems* 23.
- ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2008). Time varying undirected graphs. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT'08)*.
- ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2009). Compressed and privacy sensitive sparse regression. *IEEE Transactions on Information Theory* **55** 846–866.
- ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research* **12** 2975–3026.

## A Estimation of the covariance matrices

Theorem A.1 and A.2 show the rates of convergence in the operator and the Frobenius norm for estimating both  $A_0 \otimes B_0$  and its inverse, which will depend on  $n$  through  $\lambda_{A_0}$  and  $\lambda_{B_0}$ , which in turn depend on the rates of convergence in entry-wise max norm in estimating  $\rho(A_0)$  and  $\rho(B_0)$  respectively.

**Theorem A.1.** Let  $\lambda_{A_0} \wedge \lambda_{B_0} := \min\{\lambda_{A_0}, \lambda_{B_0}\}$ . Suppose (A1) and (A2) hold. Let us define for  $c = \frac{\log n}{\log \log(m \vee f)}$ ,  $d = 1 - (c/2 \wedge 1/2)$ . Suppose that  $\lambda_{A_0}, \lambda_{B_0} < 1$ , and for some  $0 < \varepsilon_1, \varepsilon_2 < 1$

$$\lambda_{A_0} := 3\alpha_n/\varepsilon_1 \text{ and } \lambda_{B_0} := 3\beta_n/\varepsilon_2$$

where  $\alpha_n$  and  $\beta_n$  are as defined in Theorem 4.1. Then on event  $\mathcal{X}_0$ , where  $\mathbb{P}(\mathcal{X}_0) \geq 1 - \frac{3}{\max(m,f)^2}$ , we have for  $9 < C, C' < 18$ ,

$$\begin{aligned} \left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_2 &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|A_0\|_2 \|B_0\|_2 \\ &+ 2C \lambda_{B_0} a_{\max} \|B_0\|_2 \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \\ &+ 2C' \lambda_{A_0} b_{\max} \|A_0\|_2 \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} \\ &+ 7CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}, \end{aligned}$$

and for  $19/2 > C, C' > 5$ ,

$$\begin{aligned} \left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\|_2 &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} \|A_0^{-1}\|_2 \|B_0^{-1}\|_2 \\ &+ \lambda_{B_0} \|B_0^{-1}\|_2 \frac{2C \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} \varphi_{\min}^2(\rho(A_0))} + \lambda_{A_0} \|A_0^{-1}\|_2 \frac{2C' \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B_0))} \\ &+ \frac{6CC' \lambda_{A_0} \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} b_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}^2(\rho(B_0))}. \end{aligned}$$

**Theorem A.2.** Suppose all conditions in Theorem A.1 hold. Suppose (A1) and (A2) hold. Then on event  $\mathcal{X}_0$ , we have for  $9 < C, C' < 18$ ,

$$\begin{aligned} \left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|A_0 \otimes B_0\|_F \\ &+ 2C \lambda_{B_0} a_{\max} \|B_0\|_F \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} \\ &+ 2C' \lambda_{A_0} b_{\max} \|A_0\|_F \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f} \\ &+ 7CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}, \end{aligned}$$

and for  $19/2 > C, C' > 5$ ,

$$\begin{aligned} \left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\|_F &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} \|A_0^{-1}\|_F \|B_0^{-1}\|_F \\ &+ \lambda_{B_0} \|B_0^{-1}\|_F \frac{2C \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}}{a_{\min} \varphi_{\min}^2(\rho(A_0))} + \lambda_{A_0} \|A_0^{-1}\|_F \frac{2C' \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}}{b_{\min} \varphi_{\min}^2(\rho(B_0))} \\ &+ \frac{28CC' \lambda_{A_0} \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}}{5a_{\min} b_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}^2(\rho(B_0))}. \end{aligned}$$

We show an outline for proving Theorems A.1 and A.2 in Section E, with actual proof in Section F.1 and F.2.

Corollary A.3 states the rates of convergence in the operator and the Frobenius norm for estimating  $A_*$  and  $B_*$ , where recall that  $a_{*,\max} = \max_i a_{*,ii}$  and  $a_{*,\min} = \min_i a_{*,ii}$ . Denote by  $b_{*,\max} = \max_i b_{*,ii}$  and  $b_{*,\min} = \min_i b_{*,ii}$ .



**Corollary A.3.** Let  $A_*, B_*$  be as defined in (16) and Let  $\hat{A}_*, \hat{B}_*$  be as defined in (17). Suppose Assumptions (A1) and (A2) hold. We have for some absolute constants  $9 < C, C' < 18$ , on event  $\mathcal{X}_0$ ,

$$\begin{aligned}\|\hat{A}_* - A_*\|_2 &\leq 4\frac{1}{3}C\lambda_{B_0}a_{*,\max}\kappa(\rho(A_0))^2\sqrt{|A_0^{-1}|_{0,\text{off}}}\vee 1, \\ \|\hat{B}_* - B_*\|_2 &\leq 2C'\lambda_{A_0}b_{*,\max}\kappa(\rho(B_0))^2\sqrt{|B_0^{-1}|_{0,\text{off}}}\vee 1,\end{aligned}$$

For the inverses, we have for  $5.25 < C, C' < 19/2$ , on event  $\mathcal{X}_0$ ,

$$\begin{aligned}\|\hat{A}_*^{-1} - A_*^{-1}\|_2 &\leq 4C\lambda_{B_0}\sqrt{|A_0^{-1}|_{0,\text{off}}}\vee 1/(a_{*,\min}\varphi_{\min}^2(\rho(A_0))) \\ \text{and } \|\hat{B}_*^{-1} - B_*^{-1}\|_2 &\leq 2C'\lambda_{A_0}\sqrt{|B_0^{-1}|_{0,\text{off}}}\vee 1/(b_{*,\min}\varphi_{\min}^2(\rho(B_0)))\end{aligned}$$

We have the following rate of convergence for the Frobenius norm: for  $9 < C, C' < 18$ , on event  $\mathcal{X}_0$ ,

$$\begin{aligned}\|\hat{A}_* - A_*\|_F &\leq 4\frac{1}{3}C\lambda_{B_0}a_{*,\max}\kappa(\rho(A_0))^2\sqrt{|A_0^{-1}|_{0,\text{off}}}\vee m, \\ \|\hat{B}_* - B_*\|_F &\leq 2C'\lambda_{A_0}b_{*,\max}\kappa(\rho(B_0))^2\sqrt{|B_0^{-1}|_{0,\text{off}}}\vee f\end{aligned}$$

and for  $5.25 < C, C' < 19/2$ , on event  $\mathcal{X}_0$ ,

$$\begin{aligned}\|\hat{A}_*^{-1} - A_*^{-1}\|_F &\leq 4C\lambda_{B_0}\sqrt{|A_0^{-1}|_{0,\text{off}}}\vee m/(a_{*,\min}\varphi_{\min}^2(\rho(A_0))) \\ \text{and } \|\hat{B}_*^{-1} - B_*^{-1}\|_F &\leq 2C'\lambda_{A_0}\sqrt{|B_0^{-1}|_{0,\text{off}}}\vee f/(b_{*,\min}\varphi_{\min}^2(\rho(B_0))).\end{aligned}$$

## A.1 Some convenient bounds

Throughout our proofs, we need the following bounds:

$$\frac{1}{\varphi_{\min}(A_0)} = \|A_0^{-1}\|_2 \leq \frac{\|\rho(A_0)^{-1}\|_2}{a_{\min}} = \frac{1}{a_{\min}\varphi_{\min}(\rho(A_0))}, \quad (42a)$$

$$\frac{1}{\varphi_{\min}(B_0)} = \|B_0^{-1}\|_2 \leq \frac{\|\rho(B_0)^{-1}\|_2}{b_{\min}} = \frac{1}{b_{\min}\varphi_{\min}(\rho(B_0))}, \quad (42b)$$

$$\frac{1}{\varphi_{\min}(\rho(A_0))} = \|\rho(A_0)^{-1}\|_2 \leq a_{\max}\|A_0^{-1}\|_2, \quad (42c)$$

$$\frac{1}{\varphi_{\min}(\rho(B_0))} = \|\rho(B_0)^{-1}\|_2 \leq b_{\max}\|B_0^{-1}\|_2, \quad (42d)$$

$$\|A_0\|_2 \leq a_{\max}\|\rho(A_0)\|_2 \quad \text{and} \quad \|B_0\|_2 \leq b_{\max}\|\rho(B_0)\|_2, \quad (42e)$$

$$\|\rho(A_0)\|_2 \leq \frac{\|A_0\|_2}{a_{\min}} \quad \text{and} \quad \|\rho(B_0)\|_2 \leq \frac{\|B_0\|_2}{b_{\min}}. \quad (42f)$$

## A.2 Proof of Theorem 3.1

Theorem 3.1 is a simple corollary of Theorem A.1. We now insert the bounds as in (42e) in Theorem A.1 to obtain  $\|\widehat{A \otimes B} - A_0 \otimes B_0\|_2 \leq \|A_0\|_2 \|B_0\|_2 \delta$  where

$$\begin{aligned} \delta &= \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} + 2C \frac{a_{\max} \kappa(\rho(A_0))}{a_{\min} \varphi_{\min}(\rho(A_0))} \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \\ &+ 2C' \frac{\kappa(\rho(B_0)) b_{\max}}{b_{\min} \varphi_{\min}(\rho(B_0))} \lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} \\ &+ 7CC' \lambda_{A_0} \lambda_{B_0} \frac{a_{\max} b_{\max} \kappa(\rho(A_0)) \kappa(\rho(B_0))}{a_{\min} b_{\min}} \frac{\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}(\rho(A_0)) \varphi_{\min}(\rho(B_0))} \\ &\asymp \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} + \log^d \max(m, f) \left( \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee 1}{nf}} + \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee 1}{nm}} \right) + o(1). \end{aligned}$$

For the inverse, we plug in the bounds as in (42c) and (42d) in Theorem A.1 to obtain  $\|\widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1}\|_2 \leq \|B_0^{-1}\|_2 \|A_0^{-1}\|_2 \delta'$  where

$$\begin{aligned} \delta' &= \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} + \frac{\lambda_{B_0} 2C a_{\max} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} \varphi_{\min}(\rho(A_0))} + \frac{2C' \lambda_{A_0} b_{\max} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}(\rho(B_0))} \\ &+ \frac{6CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} b_{\min} \varphi_{\min}(\rho(A_0)) \varphi_{\min}(\rho(B_0))} \\ &\asymp \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} + \log^d \max(m, f) \left( \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee 1}{nf}} + \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee 1}{nm}} \right) + o(1). \quad \square \end{aligned}$$

## A.3 Proof of Theorem 3.2

First we state the following bounds.

$$(a_{\min} \vee \varphi_{\min}(A_0)) \sqrt{m} \leq \|A_0\|_F = \sqrt{\sum_{i=1}^m \varphi_i(A_0)^2} \leq \sqrt{m} \|A_0\|_2, \quad (43a)$$

$$(b_{\min} \vee \varphi_{\min}(B_0)) \sqrt{f} \leq \|B_0\|_F = \sqrt{\sum_{i=1}^f \varphi_i(B_0)^2} \leq \sqrt{f} \|B_0\|_2, \quad (43b)$$

$$\sqrt{m}/a_{\max} = \left( \frac{1}{a_{\max}} \vee \frac{1}{\varphi_{\max}(A_0)} \right) \sqrt{m} \leq \|A_0^{-1}\|_F \leq \sqrt{m} \|A_0^{-1}\|_2, \quad (43c)$$

$$\sqrt{f}/b_{\max} = \left( \frac{1}{b_{\max}} \vee \frac{1}{\varphi_{\max}(B_0)} \right) \sqrt{f} \leq \|B_0^{-1}\|_F \leq \sqrt{f} \|B_0^{-1}\|_2. \quad (43d)$$

We now insert the lower bounds as in (43c) and (43d) in Theorem A.2 to obtain  $\left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\|_F \leq \|A_0^{-1}\|_F \|B_0^{-1}\|_F \delta'$  where

$$\begin{aligned}
\delta' &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} + \frac{2C\lambda_{B_0}\sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}}{a_{\min} \|A_0^{-1}\|_F \varphi_{\min}^2(\rho(A_0))} + \frac{2C'\lambda_{A_0}\sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}}{b_{\min} \|B_0^{-1}\|_F \varphi_{\min}^2(\rho(B_0))} + \\
&\quad + \frac{28\sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}\sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}CC'\lambda_{A_0}\lambda_{B_0}}{5a_{\min}b_{\min}\varphi_{\min}^2(\rho(A_0))\varphi_{\min}^2(\rho(B_0))\|A_0^{-1}\|_F\|B_0^{-1}\|_F} \\
&\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} + \frac{2Ca_{\max}\lambda_{B_0}}{a_{\min}\varphi_{\min}^2(\rho(A_0))}\sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} \\
&\quad + \frac{2C'b_{\max}\lambda_{A_0}}{b_{\min}\varphi_{\min}^2(\rho(B_0))}\sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}} \\
&\quad + \frac{28a_{\max}b_{\max}CC'\lambda_{A_0}\lambda_{B_0}}{5a_{\min}b_{\min}\varphi_{\min}^2(\rho(A_0))\varphi_{\min}^2(\rho(B_0))}\sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}}\sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}}.
\end{aligned}$$

Similarly, plug in the lower bounds as in (43a) and (43b) in Theorem A.2 to obtain under (A1) and (A2),  $\left\| \widehat{A \otimes B} - A_0 \otimes B_0 \right\|_F \leq \|A_0\|_F \|B_0\|_F \delta$  where

$$\begin{aligned}
\delta &\leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} + \frac{2C\kappa(\rho(A_0))^2 a_{\max}}{a_{\min}} \lambda_{B_0} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} \\
&\quad + \frac{2C'\kappa(\rho(B_0))^2 b_{\max}}{b_{\min}} \lambda_{A_0} \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}} \\
&\quad + 7CC'\lambda_{A_0}\lambda_{B_0} \frac{a_{\max}b_{\max}\kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2}{a_{\min}b_{\min}} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}} \\
&= O\left(\lambda_{B_0} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} + \lambda_{A_0} \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}}\right) \rightarrow 0.
\end{aligned}$$

To see the last equality:

1. If  $|A_0^{-1}|_{0,\text{off}} \leq m$ , then  $\lambda_{B_0} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} = \lambda_{B_0}$ ; otherwise,  $\lambda_{B_0} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee m}{m}} = o\left(\frac{1}{\sqrt{m}}\right)$ ;
2. If  $|B_0^{-1}|_{0,\text{off}} \leq f$ , then  $\lambda_{A_0} \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}} = \lambda_{A_0}$ ; otherwise,  $\lambda_{A_0} \sqrt{\frac{|B_0^{-1}|_{0,\text{off}} \vee f}{f}} = o\left(\frac{1}{\sqrt{f}}\right)$ .

Moreover, the expression for  $\delta$  can be simplified as follows:

1. If  $|A_0^{-1}|_{0,\text{off}} \leq m$  or  $\asymp m$ , and  $|B_0^{-1}|_{0,\text{off}} \leq f$  or  $\asymp f$ , then  $\delta \asymp \lambda_{B_0} + \lambda_{A_0}$ .

2. If  $1 \leq n \leq \log(m \vee f)$ , then

$$\begin{aligned}\delta &= O\left(\frac{1}{\sqrt{f}} + \frac{1}{\sqrt{m}} + \lambda_{A_0} + \lambda_{B_0}\right) \\ &\asymp \log^d \max(m, f) \left(\frac{1}{\sqrt{nf}} + \frac{1}{\sqrt{nm}}\right) \asymp \lambda_{A_0} + \lambda_{B_0}\end{aligned}$$

where  $d = 1$  for  $n = 1$ , and  $d \geq 1/2$  for  $n \leq \log(m \vee f)$ .

The upper bounds on  $\delta'$  are dominated by that of  $\delta$  under each case; hence the same conclusions hold.  $\square$

## B Concentration inequalities for the Gaussian Chaos

In order to prove Lemma 4.2, and Theorem C.1, we need the following large deviation bound on the sum of i.i.d. random variables which are Gaussian chaos of order 2. The proof of Theorem B.1 appears in Section B.1.

**Theorem B.1.** *Let  $(g_1, \dots, g_m)$  be a random vector with  $m$  independent  $N(0, 1)$  variables and  $Y = \sum_{i=1}^m \sum_{j=1}^m A_{i,j} g_i g_j$ . Let*

$$Z := \frac{Y - \mathbb{E}[Y]}{\sqrt{m}} = \frac{1}{\sqrt{m}} \left( \sum_{k=1}^m \sum_{j=1, j \neq k}^m g_k g_j A_{kj} + \sum_{k=1}^m (g_k^2 - 1) A_{kk} \right),$$

and  $Z_1, \dots, Z_n$  be independent copies of  $Z$ . Then Bernstein's inequality yields

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{\ell=1}^n Z_\ell \right| \geq \tau \right) \leq 2 \exp \left( -\frac{n\tau^2}{2v_1 + 2W\tau} \right),$$

and an improved bound based on the Bennet's inequality yields the following:

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n |Z_i| \right| \geq \tau \right) \leq 2 \exp \left( -\frac{n\tau^2}{v_1 + W\tau + v_1 \sqrt{1 + 2\frac{W}{v_1}\tau}} \right) \quad (44)$$

where  $v_1 = \frac{C_1}{2} \frac{\|A\|_F^2}{m} < \infty$  and  $W = \frac{C_2}{2} \frac{\|A\|_F}{m} < \infty$  with  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$  and  $C_2 = \sqrt{8e} \approx 7.6885$ .

### B.1 Proof of Theorem B.1

This proof follows exactly the sequence of arguments in Zhou et al. (2009). Hence we only sketch it here for completeness. Applying a general bound of Ledoux and Talagrand (1991) for Gaussian chaos gives that

$$\mathbb{E} [|Z|^r] \leq (r-1)^r (\mathbb{E} [|Z|^2])^{r/2} \quad (45)$$

for all  $r > 2$ . This bound guaranties that factorial-type moment estimate for Gaussian chaos, which makes application of Bennet's inequality possible. We first state the following claim derived from (45), whose proof appears in Rauhut et al. (2008), which we omit.

**Claim B.2.** (**Rauhut et al. (2008)**) Let  $W = e(\mathbb{E}[|Z|^2])^{1/2}$  and  $s = \frac{2e}{\sqrt{6\pi}}\mathbb{E}[|Z|^2]$ .

$$\forall r > 2, \quad \mathbb{E}[|Z|^r] \leq r!W^{r-2}s/2.$$

Clearly the above claim holds for  $r = 2$ , since trivially  $\mathbb{E}[|Z|^r] \leq r!W^{r-2}s/2$  given that for  $r = 2$

$$r!W^{r-2}s/2 = 2W^{2-2}s/2 = s = \frac{2e}{\sqrt{6\pi}}\mathbb{E}[|Z|^2] \approx 1.2522\mathbb{E}[|Z|^2]$$

where

$$\mathbb{E}[|Z|^2] = \frac{1}{m} \left( \sum_{k \neq j} A_{j,k}^2 + 2 \sum_{k=1}^m A_{kk}^2 \right) = \frac{1}{m} (\|A\|_F^2 + \|\text{diag } A\|_F^2) \leq \infty$$

Set

$$W = e(\mathbb{E}[|Z|^2])^{1/2} \leq \frac{e\sqrt{2}}{\sqrt{m}} \|A\|_F < \infty$$

where the last inequality holds by assumption, and for all  $i$ ,

$$v_i = \frac{2e}{\sqrt{6\pi}}\mathbb{E}[|Z|^2] \leq \frac{4e}{\sqrt{6\pi}} \frac{1}{m} \|A\|_F^2 \approx \frac{2.5044}{m} \|A\|_F^2 < \infty.$$

Thus for independent random variables  $Z_i, \forall i = 1, \dots, n$ , we have

$$\mathbb{E}[|Z_i|^r] \leq r!W^{r-2}v_i/2.$$

We now apply the following Theorem B.3, the proof of which follows arguments from expression (1) to (7) in **Bennett (1962)**, where one needs to replace  $\sigma^2$  with  $v$  and  $\sigma_i^2$  with  $v_i$  everywhere in between (1) to (7) while noticing that all arguments will necessarily go through despite these substitutions.

**Theorem B.3. (Bernstein's and Bennet's inequalities **Bennett (1962)**)** Let  $Z_1, \dots, Z_n$  be independent random variables with zero mean such that

$$\mathbb{E}[|Z_i|^r] \leq r!W^{r-2}v_i/2,$$

for every  $r \geq 2$  and some constant  $W$  and  $v_i, \forall i = 1, \dots, n$ . Then for  $\tau > 0$ ,

$$\mathbb{P} \left( \left| \sum_{i=1}^n |Z_i| \right| \geq \tau \right) \leq 2 \exp \left( -\frac{\tau^2}{2v + 2W\tau} \right)$$

with  $v = \sum_{i=1}^n v_i$ ; An improved bound is also given by

$$\mathbb{P} \left( \left| \sum_{i=1}^n |Z_i| \right| \geq \tau \right) \leq 2 \exp \left( -\frac{\tau^2}{v + W\tau + \sqrt{v^2 + 2Wv\tau}} \right) \quad (46)$$

**Remark B.4.** The statement in Theorem B.3 corresponds to Expression (7) in **Bennett (1962)**, once we replace  $\sigma^2$  by  $v = \sum_{i=1}^n v_i$ , which is typically known as the Bernstein's inequality; Similarly, (46) corresponds to Expression (7a) in **Bennett (1962)**.

We can then apply the Bernstein's Inequality to obtain the following:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^n Z_{\ell}\right| \geq \tau\right) &= \mathbb{P}\left(\left|\sum_{\ell=1}^n Z_{\ell}\right| \geq n\tau\right) \\
&\leq 2\exp\left(-\frac{(n\tau)^2}{2\sum_{i=1}^n v_i + 2Wn\tau}\right) = 2\exp\left(-\frac{n\tau^2}{2v_1 + 2W\tau}\right) \\
&\leq 2\exp\left(-\frac{n\tau^2}{C_1(\|A\|_F^2/m) + C_2(\|A\|_F/\sqrt{m})\tau}\right)
\end{aligned}$$

with  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$  and  $C_2 = \sqrt{8e} \approx 7.6885$  as desired. Similarly, we apply the Bennet's inequality in (46) to obtain (44).  $\square$

## B.2 Proof of Theorem 4.1

Throughout this proof, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. First we obtain the large deviation bounds on the estimated correlation coefficients. We prove it only for  $\hat{\Gamma}(B_0)$ , as a similar argument also works for  $\hat{\Gamma}(A_0)$ . On  $\mathcal{X}_0$ , we have

$$\begin{aligned}
&\forall i, \quad \left| \frac{\frac{1}{n}\sum_{t=1}^n \|y(t)^i\|_2^2}{b_{ii}\text{tr}(A_0)} - 1 \right| \leq \frac{m}{\text{tr}(A_0)}\tau_0 = \alpha_n, \\
&\text{and hence} \quad \frac{\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^i\|_2^2}}{\sqrt{b_{ii}\text{tr}(A_0)}} \geq \sqrt{1 - \alpha_n},
\end{aligned}$$

$$\text{and } \forall i \neq j, \quad \left| \frac{\frac{1}{n}\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\text{tr}(A_0)\sqrt{b_{ii}, b_{jj}}} - \frac{b_{ij}}{\sqrt{b_{ii}, b_{jj}}} \right| \leq \frac{m}{\text{tr}(A_0)}\tau_0.$$

For all  $i, j$ , and  $\alpha_n := \frac{m}{\text{tr}(A_0)}\tau_0 < 1/3$ , we have on  $\mathcal{X}_0$ ,

$$\begin{aligned}
|\hat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| &:= \left| \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{\sum_{t=1}^n \|y(t)^i\|_2^2} \sqrt{\sum_{t=1}^n \|y(t)^j\|_2^2}} - \rho_{i,j}(B_0) \right| \\
&= \left| \frac{\frac{1}{n}\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle / (\text{tr}(A_0)\sqrt{b_{ii}b_{jj}})}{(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^i\|_2^2 / (b_{ii}\text{tr}(A_0))})(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^j\|_2^2 / (b_{jj}\text{tr}(A_0))})} - \rho_{i,j}(B_0) \right| \\
&= \left| \frac{\frac{1}{n}\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle / (\text{tr}(A_0)\sqrt{b_{ii}b_{jj}}) - \rho_{i,j}(B_0)}{(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^i\|_2^2 / (b_{ii}\text{tr}(A_0))})(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^j\|_2^2 / (b_{jj}\text{tr}(A_0))})} \right| \\
&+ \left| \frac{\rho_{i,j}(B_0)}{(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^i\|_2^2 / (b_{ii}\text{tr}(A_0))})(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^j\|_2^2 / (b_{jj}\text{tr}(A_0))})} - \rho_{i,j}(B_0) \right| \\
&\leq \frac{\alpha_n}{1 - \alpha_n} + |\rho_{i,j}(B_0)| \left| \frac{1}{1 - \alpha_n} - 1 \right| \leq 3\alpha_n = \frac{3m}{\text{tr}(A_0)}\tau_0.
\end{aligned}$$

The last inequality in the theorem statement follows immediately by summing over the large deviation inequalities on the  $\ell_2^2$  norms for the row vectors and column vectors respectively. In more details, we have

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 - \text{tr}(A_0)\text{tr}(B_0) \right| = \left| \frac{1}{n} \sum_{t=1}^n \sum_{i=1}^m \|x(t)\|_2^2 - \text{tr}(A_0)\text{tr}(B_0) \right| \\
&= \left| \frac{1}{n} \sum_{t=1}^n \sum_{j=1}^f \|y(t)\|_2^2 - \text{tr}(A_0)\text{tr}(B_0) \right| \\
&\leq \min \left\{ \sum_{i=1}^m \left| \frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2 - a_{ii}\text{tr}(B_0) \right|, \sum_{j=1}^f \left| \frac{1}{n} \sum_{t=1}^n \|y(t)^j\|_2^2 - b_{jj}\text{tr}(A_0) \right| \right\} \\
&\leq \min \{ \text{tr}(A_0)f\tau'_0, \text{tr}(B_0)m\tau_0 \} = (\alpha_n \wedge \beta_n)\text{tr}(A_0)\text{tr}(B_0).
\end{aligned}$$

The theorem is thus proved.  $\square$

### B.3 Proof of Lemma 4.2

The conclusions of Lemma 4.2 follow from Theorem C.1. In proving Theorem C.1, we will focus on the case when  $n > \log(m \vee f)$ . Hence we focus on the case when  $n \leq \log(m \vee f)$  in the proof of Lemma 4.2.

Let us first define a set of mean-zero random variables  $Z_A, Z_B$ , and  $Z(i, j), \forall i, j \in \{1, \dots, m\}, i \neq j$ . For two random variables  $Y, Z$ ,  $Y \sim Z$  means that they follow the same distribution.

**Proposition B.5.** *Let  $g_1, g_2, \dots \sim N(0, 1)$ . We have for row vectors  $y^1, \dots, y^f$  and column vectors  $x^1, \dots, x^m$  of  $X \sim \mathcal{N}_{f,m}(0, A_0 \otimes B_0)$ , where  $A_0 = (a_{ij})$  and  $B_0 = (b_{ij})$ ,*

$$\frac{(\|y^j\|_2^2/b_{jj}) - \text{tr}(A_0)}{\sqrt{m}} \sim Z_A \quad (47)$$

$$\begin{aligned}
& \text{where } Z_A := \frac{1}{\sqrt{m}} \left( \sum_{i=1}^m \sum_{j=1}^m a_{i,j} g_i g_j - \text{tr}(A_0) \right), \\
& \frac{(\|x^j\|_2^2/a_{jj}) - \text{tr}(B_0)}{\sqrt{f}} \sim Z_B \quad (48)
\end{aligned}$$

$$\text{where } Z_B := \frac{1}{\sqrt{f}} \left( \sum_{i=1}^f \sum_{j=1}^f b_{i,j} g_i g_j - \text{tr}(B_0) \right).$$

Let  $\begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}$  be the unique symmetric square root of the positive definite matrix  $B_{0,\{i,j\}}$ , where  $B_{0,\{i,j\}}$  is the submatrix of  $B_0$  with rows and columns indexed by  $\{i, j\}$ . Define

$$Z(i, j) := \frac{1}{\sqrt{2m}} \left( \sum_{k=1}^{2m} \sum_{\ell=1}^{2m} \frac{M'_{k\ell} g_k g_\ell}{\sqrt{b_{ii}b_{jj}}} - \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \text{tr}(A_0) \right) \quad (49)$$

$$\text{where } M' = \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix} \otimes A_0 \text{ and } \|M'\|_F^2 = b_{ii}b_{jj} \|A_0\|_F^2.$$

Then  $\forall i, j$

$$\frac{\langle y^i, y^j \rangle - b_{ij}\text{tr}(A_0)}{\sqrt{2m}\sqrt{b_{ii}b_{jj}}} \sim Z(i, j). \quad (50)$$



We assume that  $\max(m, f) \geq 2$ . We first apply the Bennett's Inequality as stated in (44) in Theorem B.1 with  $n = (\log \max(m, f))^c$ , where  $1 \geq c \geq 0$ , and  $\varepsilon = \sqrt{m}\tau_0 = 20 \frac{\|A_0\|_F}{\sqrt{m}} \log^{d-c/2}(m \vee f)$ , where  $d = 1 - c/2 \geq c/2$ , to obtain a concentration bound for  $\sum_{t=1}^n Z_t$ , where  $Z_1, \dots, Z_n$  are independent copies of  $Z_B$  as defined in (48). Assuming that  $\max(m, f) \geq 2$ , and hence  $\log^{d-c/2}(m \vee f) \geq 1$ , we have

$$\begin{aligned}
& \mathbb{P} \left( \exists i : \left| \frac{1}{m} \left| \frac{1}{n} \sum_{t=1}^n \frac{\langle y(t)^i, y(t)^i \rangle}{b_{ii}} - \text{tr}(A_0) \right| \geq \tau_0 \right) \leq f \mathbb{P} \left( \frac{1}{\sqrt{m}} \left| \frac{1}{n} \sum_{t=1}^n Z_t \right| \geq \tau_0 \right) \\
&= f \mathbb{P} \left( \left| \sum_{\ell=1}^n Z_\ell \right| \geq n\varepsilon \right) \leq 2f \exp \left( - \frac{n\varepsilon^2}{v_1 + W\varepsilon + v_1 \sqrt{1 + \frac{2W}{v_1} \varepsilon}} \right) \\
&\leq 2f \exp \left( - \frac{20^2 \log^{2d}(m \vee f)}{2.5044 + 3.8443 \cdot 20 \log^{d-c/2}(m \vee f) + 2.5044 \sqrt{1 + \frac{7.6885}{2.5044} 20 \log^{d-c/2}(m \vee f)}} \right) \\
&\leq 2f \exp \left( - \frac{20^2 \log^{2d}(m \vee f)}{\left( 2.5044 + 3.8443 \cdot 20 + 2.5044 \sqrt{1 + \frac{7.6885}{2.5044} 20} \right) (\log(m \vee f))^{d-c/2}} \right) \\
&\leq \frac{2f}{\max(m, f)^4}
\end{aligned}$$

where  $v_i = \frac{4e}{\sqrt{6\pi}} \frac{\|A_0\|_F^2}{m} \approx 2.5044 \frac{\|A_0\|_F^2}{m}$  for all  $i$ , and  $W = \sqrt{2}e \frac{\|A_0\|_F}{\sqrt{m}} \approx 3.8443 \frac{\|A_0\|_F}{\sqrt{m}}$ . Let

$$Z_t(i, j) := \frac{\langle y(t)^i, y(t)^j \rangle - b_{ij} \text{tr}(A_0)}{\sqrt{b_{ii} b_{jj}} \sqrt{2m}}.$$

Then  $Z_1(i, j), \dots, Z_n(i, j)$  are independent copies of  $Z(i, j)$  as defined in (49). We next apply the Bennett's Inequality as stated in Theorem B.1 with  $n = (\log \max(m, f))^c$ , where  $0 \leq c \leq 1$ , and  $\varepsilon = \sqrt{m/2}\tau_0 = 20 \frac{1}{\sqrt{2}} \frac{\|A_0\|_F}{\sqrt{m}} \frac{(\log \max(m, f))^d}{\sqrt{n}}$ . By definition of  $Z_t(i, j)$  as in (49) and expression (50), we have

$$\begin{aligned}
& \mathbb{P} \left( \exists i \neq j : \frac{1}{m} \left| \frac{1}{n} \sum_{t=1}^n \frac{\langle y(t)^i, y(t)^j \rangle}{\sqrt{b_{ii} b_{jj}}} - \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}} \text{tr}(A_0) \right| \geq \tau_0 \right) \\
&\leq \frac{f(f-1)}{2} \mathbb{P} \left( \left| \frac{1}{n} \sum_{t=1}^n Z_t(i, j) \right| \geq \sqrt{\frac{m}{2}} \tau_0 \right) = \frac{f(f-1)}{2} \mathbb{P} \left( \left| \sum_{t=1}^n Z_t(i, j) \right| \geq n\varepsilon \right) \\
&\leq f(f-1) \exp \left( - \frac{(n\varepsilon)^2}{\sum_{i=1}^n v_i + Wn\varepsilon + v \sqrt{1 + \frac{2W}{v} n\varepsilon}} \right) \\
&= f(f-1) \exp \left( - \frac{n\varepsilon^2}{v_1 + W\varepsilon + v_1 \sqrt{1 + \frac{2W}{v_1} \varepsilon}} \right) \\
&\leq f(f-1) \exp \left( - \frac{20^2 \log^{2d}(m \vee f)}{2.5044 + 3.8443 \cdot 20 (\log(m \vee f))^{d-c/2} + 2.5044 \sqrt{1 + \frac{7.6885}{2.5044} 20 (\log(m \vee f))^{d-c/2}}} \right) \\
&\leq f(f-1) \exp \left( -4 (\log \max(m, f))^{d+c/2} \right) \leq \frac{f(f-1)}{\max(m, f)^4},
\end{aligned}$$

where  $v_1 = 2.5044 \frac{1}{2m} \|A_0\|_F^2$ , and  $W = \frac{e\sqrt{2}}{\sqrt{2m}} \|A_0\|_F$ . Following similar arguments, we get, assuming that  $\max(m, f) \geq 2$ ,

$$\begin{aligned} \mathbb{P}\left(\exists i : \frac{1}{f} \left| \frac{1}{n} \sum_{t=1}^n \frac{\|x(t)^i\|_2^2}{a_{ii}} - \text{tr}(B_0) \right| \geq \tau'_0\right) &\leq \frac{2m}{\max(m, f)^4}, \\ \text{and } \mathbb{P}\left(\exists i \neq j : \frac{1}{f} \left| \frac{1}{n} \sum_{t=1}^n \frac{\langle x(t)^i, x(t)^j \rangle}{\sqrt{a_{ii}a_{jj}}} - \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}} \text{tr}(B_0) \right| \geq \tau'_0\right) &\leq \frac{m(m-1)}{\max(m, f)^4}. \end{aligned}$$

The lemma is thus proved by summing up the probability for all bad events and using the fact that for  $(m \vee f) \geq 2$ ,

$$\begin{aligned} &\frac{m(m-1) + 2m + f(f-1) + 2f}{\max(m, f)^4} \\ &\leq \frac{2 \max(m, f)^2 + 2(m \vee f)}{\max(m, f)^4} \leq \frac{3 \max(m, f)^2}{\max(m, f)^4}. \end{aligned}$$

□

It remains to prove Proposition B.5.

*Proof of Proposition B.5.* First observe that when  $i = j$ ,

$$\|y^j\|_2^2 = \sum_{k=1}^m x_j^k \times x_j^k = (g_1, \dots, g_m) b_{jj} A_0 (g_1, \dots, g_m)^T = b_{jj} \sum_{k=1}^m \sum_{\ell=1}^m a_{k\ell} g_k g_\ell,$$

where  $g_1, \dots, g_m \sim N(0, 1)$  and the mean-zero normal random vector  $y^j \sim \mathcal{N}_m(0, b_{jj} A_0)$  can be expressed as:

$$y^j = (x_j^k)_{k=1}^m = (A_0 b_{jj})^{1/2} (g_1, \dots, g_m)^T.$$

Statements in (48) and (47) follow immediately. The rest of the proof follows exactly that of Theorem C.1, in the decorrelation step, for  $M = I$ , after we write  $\langle y^1, y^2 \rangle = \sum_{k=1}^m x_1^k \times x_2^k = \sum_{k=1}^m \sum_{\ell=1}^m I_{k\ell} x_1^\ell \times x_2^k$ .

□

## C A general theory on concentration inequalities

Suppose that we have  $n$  i.i.d. random matrices  $X(1), X(2), \dots, X(n) \sim \mathcal{N}_{f,m}(0, A_0 \otimes B_0)$ , where  $A_0 = (a_{jk})$  and  $B_0 = (b_{jk})$  are positive definite. Then

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n x(t)^\ell \otimes x(t)^k &= \hat{S}_n^{\ell k} \quad \text{where } \mathbb{E} \left[ x(t)^\ell \otimes x(t)^k \right] = a_{\ell k} B_0, \\ \frac{1}{n} \sum_{t=1}^n y(t)^\ell \otimes y(t)^k &= \tilde{S}_n^{\ell k} \quad \text{where } \mathbb{E} \left[ y(t)^\ell \otimes y(t)^k \right] = b_{\ell k} A_0, \end{aligned}$$

where  $x(t)^1, \dots, x(t)^m \in \mathbf{R}^f$  are column vectors, and  $y(t)^1, \dots, y(t)^f \in \mathbf{R}^m$  are row vectors of matrix  $X(t)$ . We now provide a large deviation inequality for the weighted sum of submatrices  $\hat{S}_n^{\ell k}, \ell, k = 1, \dots, m$  (or sum of  $\tilde{S}_n^{\ell k}, \ell, k = 1, \dots, f$ ), where the weights correspond to entries in an  $m \times m$  matrix  $M$  (or in an  $f \times f$  matrix  $N$ ). More precisely, we prove Theorem C.1.

**Theorem C.1.** Let  $m \vee f \geq 2$ . Let matrices  $M_{m \times m}$  and  $N_{f \times f}$  satisfy the following condition:

$$\frac{1}{m} \|M\|_F^2 < \infty \text{ and } \frac{1}{f} \|N\|_F^2 < \infty. \quad (51)$$

Denote by  $D = \|A_0^{1/2} M A_0^{1/2}\|_F / \sqrt{m}$  and  $D' = \|B_0^{1/2} N B_0^{1/2}\|_F / \sqrt{f}$ . Suppose that  $n \leq \log(m \vee f)$ . Then, we have for  $d = 1 - (c/2 \wedge 1/2)$ , where  $c = \frac{\log n}{\log \log(m \vee f)}$ ,

$$\begin{aligned} & \frac{1}{m} \left| \text{diag}(B_0)^{-1/2} \left( \sum_{k=1}^m \sum_{\ell=1}^m M_{k\ell} \widehat{S}_n^{\ell k} \right) \text{diag}(B_0)^{-1/2} - \text{tr}(A_0 M) \rho(B_0) \right|_{\max} \\ & \leq 20D \frac{\log^d \max(f, m)}{\sqrt{mn}}, \\ & \frac{1}{f} \left| \text{diag}(A_0)^{-1/2} \left( \sum_{k=1}^m \sum_{\ell=1}^m N_{k\ell} \widetilde{S}_n^{\ell k} \right) \text{diag}(A_0)^{-1/2} - \text{tr}(B_0 N) \rho(A_0) \right|_{\max} \\ & \leq 20D' \frac{\log^d \max(f, m)}{\sqrt{fn}}. \end{aligned}$$

Otherwise, let  $n > 4(C_1 \kappa^2 + C_2 \kappa) \log \max(m, f)$ , where  $\kappa > 0$ . Then, with probability  $1 - \frac{3}{\max(m, f)^2}$ , the above inequalities hold with  $d = 1/2$  and the constant 20 being replaced with  $2\sqrt{C_1 + C_2/\kappa}$ , where  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ , and  $C_2 = \sqrt{8e} \approx 7.6885$ .

**Remark C.2.** 1. Theorem C.1 is stated as a matrix max norm bound for convenience; when we look at the individual entry-wise bound, we will come to the same conclusions as in Lemma 4.2. Clearly, Theorem 4.1 and Corollary 2.2 are special cases of Theorem C.1 when we set  $M = I$ .

2. Suppose the operator norm of  $M$  is finite, then (51) holds given  $\frac{1}{m} \|M\|_F^2 \leq \|M\|_2^2$ . When we put a stronger (in terms of constants) lower bound on  $n$  by increasing  $\kappa$ , then  $C_2$  disappears from the RHS of the inequalities, and the leading constants become  $\approx 4.472$ , which is not optimized in this proof.

For the special case when  $M = A_0^{-1}$  and  $N = B_0^{-1}$ , we have the following corollary of Theorem C.1.

**Corollary C.3.** Suppose that  $n \leq \log(m \vee f)$ . Then, we have for  $d = 1 - (c/2 \wedge 1/2)$  where  $c = \frac{\log n}{\log \log(m \vee f)}$ ,

$$\begin{aligned} & \left| \frac{1}{m} \text{diag}(B_0)^{-1/2} \widetilde{B}(A_0) \text{diag}(B_0)^{-1/2} - \rho(B_0) \right|_{\max} \leq 20 \frac{\log^d \max(f, m)}{\sqrt{mn}} \\ & \left| \frac{1}{f} \text{diag}(A_0)^{-1/2} \widetilde{A}(B_0) \text{diag}(A_0)^{-1/2} - \rho(A_0) \right|_{\max} \leq 20 \frac{\log^d \max(m, f)}{\sqrt{fn}}. \end{aligned}$$

Otherwise, let  $n > 4(C_1 \kappa^2 + C_2 \kappa) \log \max(m, f)$ , where  $\kappa > 0$ . Then, with probability  $1 - \frac{3}{\max(m, f)^2}$ , the above inequalities hold with  $d = 1/2$  and the constant 20 being replaced with  $2\sqrt{C_1 + C_2/\kappa}$ , where  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ , and  $C_2 = \sqrt{8e} \approx 7.6885$ .

## C.1 Proof of Theorem C.1

Our analysis will focus on estimating  $B_0$  and  $\rho(B_0)$  when  $n > \log(m \vee f)$ , which applies to  $A_0$  and  $\rho(A_0)$  with minor changes on notation. In case  $n \leq \log(m \vee f)$ , the proof is similar to that of Lemma 4.2 in the union bound part, while for decorrelation, it follows the same arguments as below. In particular, all concentration bounds on random variables  $Z_A, Z_B$ , and  $Z_{ij}, \forall ij$  as defined and bounded in the proof of

Lemma 4.2, will apply to the corresponding ones in the current proof after we replace  $\|A_0\|_2$  with  $D$  and  $\|B_0\|_2$  with  $D'$  everywhere.

**Decorrelation.** The quantity which we are interested in is the following weighted sum of random matrices:

$$\begin{aligned} \sum_{k=1}^m \sum_{\ell=1}^m \left( \widehat{S}_n^{\ell k} M_{k\ell} - a_{\ell k} M_{k\ell} B_0 \right) = \\ \frac{1}{n} \sum_{t=1}^n \left( \sum_{k=1}^m \sum_{\ell=1}^m M(k, \ell) x(t)^\ell \otimes x(t)^k - \langle A_0, M \rangle B_0 \right). \end{aligned}$$

Let us focus on a particular entry  $b_{i,j}$  in matrix  $B_0$  and rewrite the sum above as follows,

$$\begin{aligned} \sum_{k=1}^m \sum_{\ell=1}^m \left( \widehat{S}_{i,j}^{\ell k} M_{k\ell} - a_{\ell k} M_{k\ell} b_{ij} \right) = \\ \frac{1}{n} \sum_{t=1}^n \left( \sum_{k=1}^m \sum_{\ell=1}^m M_{k\ell} x(t)_i^\ell \times x(t)_j^k - \langle A_0, M \rangle b_{ij} \right) \end{aligned}$$

where each summand  $\sum_{k=1}^m \sum_{\ell=1}^m M_{k\ell} x(t)_i^\ell \times x(t)_j^k - \langle A_0, M \rangle b_{ij}$  is a Gaussian chaos of order 2. We explore the concentration of the sum on the RHS in the next subsection. We first need to decorrelate the vectors in the sum. First observe that when  $i = j$ , the Gaussian random vector  $(x(1)_j^k)_{k=1}^m$  involved in the sum is of size  $m$ , with covariance matrix being  $b_{jj} A_0$ . Without loss of generality, we write  $(x(1)_j^k)_{k=1}^m = (A_0 b_{jj})^{1/2} (g_1, \dots, g_m)^T$ , where  $g_1, \dots, g_m \sim N(0, 1)$  and replace the sum with the following expression:

$$\begin{aligned} \sum_{k=1}^m \sum_{\ell=1}^m M_{k\ell} x(1)_j^\ell \times x(1)_j^k &= (g_1, \dots, g_m) b_{jj} A_0^{1/2} M A_0^{1/2} (g_1, \dots, g_m)^T \\ &= \sum_{k=1}^m \sum_{\ell=1}^m b_{jj} M'_{k\ell} g_k g_\ell \quad \text{where } M' = A_0^{1/2} M A_0^{1/2}. \end{aligned}$$

We next fix  $i = 1, j = 2$ . Then for vectors  $(x(1)_1^\ell)_{\ell=1}^m$  and  $(x(1)_2^k)_{k=1}^m$ , we concatenate them to become a Gaussian random vector of size  $2m$  with covariance matrix

$$\text{Cov} \left( (x(1)_1^\ell)_{\ell=1}^m, (x(1)_2^k)_{k=1}^m \right) = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \otimes A_0 := B_{0,\{1,2\}} \otimes A_0.$$

where  $B_{0,\{1,2\}}$  is the submatrix of  $B_0$  with rows and columns indexed by  $\{1, 2\}$ . Let  $\begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}$  be the unique symmetric square root of the positive definite matrix  $B_{0,\{i,j\}}$ . Thus for  $g_1, \dots, g_{2m} \sim N(0, 1)$ , we write for the concatenated vector

$$\begin{aligned} \left( (x(1)_1^\ell)_{\ell=1}^m, (x(1)_2^k)_{k=1}^m \right) &= B_{0,\{1,2\}}^{1/2} \otimes A_0^{1/2} (g_1, \dots, g_{2m})^T \\ &:= \begin{pmatrix} c_{11} A_0^{1/2} & c_{12} A_0^{1/2} \\ c_{12} A_0^{1/2} & c_{22} A_0^{1/2} \end{pmatrix} (g_1, \dots, g_{2m})^T \\ \text{and } \sum_{k=1}^m \sum_{\ell=1}^m M_{k\ell} x(1)_1^\ell \times x(1)_2^k &= (x(1)_1^1, \dots, x(1)_1^m) M (x(1)_2^1, \dots, x(1)_2^m)^T \\ &= \sum_{i=1}^{2m} \sum_{j=1}^{2m} M'_{i,j} g_i g_j \end{aligned}$$

where

$$M' = \begin{pmatrix} c_{11}c_{12} & c_{11}c_{22} \\ c_{12}c_{12} & c_{12}c_{22} \end{pmatrix} \otimes A_0^{1/2} M A_0^{1/2}$$

and  $\mathbb{E} \left[ \sum_{i=1}^{2m} \sum_{j=1}^{2m} M'_{i,j} g_i g_j \right] = \text{tr}(M') = b_{12} \text{tr}(A_0 M).$

We have  $\|M'\|_F^2 = (c_{12}^2 + c_{22}^2)(c_{11}^2 + c_{12}^2) \|A_0^{1/2} M A_0^{1/2}\|_F^2 = b_{11} b_{22} \|A_0^{1/2} M A_0^{1/2}\|_F^2.$

**Applying the union bound.** Let  $g_k, k = 1, 2, \dots,$  be independent standard Gaussian random variables. For  $j = 1, \dots, f$ , and

$$\frac{1}{\sqrt{m}} \sum_{k=1}^m \sum_{\ell=1}^m \hat{S}_{j,j}^{k\ell} M_{k\ell} / b_{jj} - \langle A_0, M \rangle = \frac{1}{n} \sum_{t=1}^n Z_t(jj)$$

where  $Z_1, \dots, Z_n$  are independent copies of mean zero random variable

$$Z_A := \frac{1}{\sqrt{m}} \left( \sum_{k=1}^m \sum_{\ell=1}^m M'_{k\ell} g_k g_\ell - \langle A_0, M \rangle \right) \quad \text{where } M' = A_0^{1/2} M A_0^{1/2}.$$

Let

$$M'(i, j) = \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix} \otimes A_0^{1/2} M A_0^{1/2}$$

where  $\|M'(i, j)\|_F = \sqrt{b_{ii}b_{jj}} \|A_0^{1/2} M A_0^{1/2}\|_F,$

and  $Z(i, j) := \frac{1}{\sqrt{2m}} \left( \sum_{k=1}^{2m} \sum_{\ell=1}^{2m} \frac{M'(i, j)_{k\ell}}{\sqrt{b_{ii}b_{jj}}} g_k g_\ell - \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \langle A_0, M \rangle \right).$

Let  $g(t)_k, k = 1, 2, \dots, t = 1, \dots, n$  be independent standard Gaussian random variables. In addition, we need to bound the sum for  $i \neq j, i, j = 1, \dots, f$ ,

$$\begin{aligned} & \frac{1}{\sqrt{2m}} \sum_{k=1}^m \sum_{\ell=1}^m \frac{\hat{S}_{i,j}^{k\ell} M_{k\ell}}{\sqrt{b_{ii}b_{jj}}} - \langle A_0, M \rangle \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \\ &= \frac{1}{\sqrt{2m}} \sum_{k=1}^m \sum_{\ell=1}^m \left( \frac{1}{n} \sum_{t=1}^n \frac{M_{k\ell} x(t)_i^\ell \times x(t)_j^k}{\sqrt{b_{ii}b_{jj}}} - \langle A_0, M \rangle \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \right) \\ &= \frac{1}{n} \sum_{t=1}^n \frac{1}{\sqrt{2m}} \left( \sum_{k=1}^{2m} \sum_{\ell=1}^{2m} \frac{M'(i, j)_{k\ell}}{\sqrt{b_{ii}b_{jj}}} g(t)_k g(t)_\ell - \langle A_0, M \rangle \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \right) \\ &= \frac{1}{n} \sum_{t=1}^n Z_t(i, j) \end{aligned}$$

where  $Z_1(i, j), \dots, Z_n(i, j)$  are independent copies of random variable  $Z_{ij}$  defined as above. We can then apply the Bernstein's Inequality to obtain the following: for  $C_1 = 5.0088$  and  $C_2 = 7.6885$ , and denote by

$$D = \left\| A_0^{1/2} M A_0^{1/2} \right\|_F / \sqrt{m}, \text{ for all } j,$$

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{t=1}^n Z_t(j, j) \right| \geq \varepsilon \right) \leq 2 \exp \left( - \frac{n\varepsilon^2}{C_1 D^2 + C_2 D \varepsilon} \right)$$

and for all  $i \neq j, i, j = 1, \dots, f, 2v_1 = C_1 D^2/2$ , and  $2W = C_2 D/\sqrt{2}$

$$\begin{aligned} \mathbb{P} \left( \left| \frac{1}{n} \sum_{t=1}^n Z_t(i, j) \right| \geq \frac{\varepsilon}{\sqrt{2}} \right) &\leq 2 \exp \left( - \frac{n\varepsilon^2/2}{2v_1 + 2W\varepsilon/\sqrt{2}} \right) \\ &\leq 2 \exp \left( - \frac{n\varepsilon^2/2}{C_1 D^2/2 + C_2 (D/\sqrt{2})\varepsilon/\sqrt{2}} \right) \\ &\leq 2 \exp \left( - \frac{n\varepsilon^2}{C_1 D^2 + C_2 D \varepsilon} \right), \end{aligned}$$

where for both inequalities, the second term in the denominator starts to dominate only when  $\varepsilon > \frac{C_1 D}{C_2}$ .

Thus we have the following by the union bound: for  $\varepsilon = c \left\| A_0^{1/2} M A_0^{1/2} \right\|_F \sqrt{\frac{\log \max(f, m)}{mn}} := cD \sqrt{\frac{\log(f \vee m)}{n}}$

$$\begin{aligned} &\mathbb{P} \left( \frac{1}{\sqrt{m}} \left| \text{diag}(B_0)^{-1/2} \sum_{k=1}^m \sum_{\ell=1}^m \left( \hat{S}_n^{k\ell} M_{k\ell} \right) \text{diag}(B_0)^{-1/2} - \langle A_0, M \rangle \rho(B_0) \right|_{\max} \geq \varepsilon \right) \\ &\leq \sum_{j=1}^f \mathbb{P} \left( \frac{1}{n} \left| \sum_{t=1}^n Z_t(j, j) \right| \geq \varepsilon \right) + \sum_{i \neq j} \mathbb{P} \left( \frac{1}{n} \left| \sum_{t=1}^n Z_t(i, j) \right| \geq \frac{\varepsilon}{\sqrt{2}} \right) \\ &\leq (f^2 + f) \exp \left( - \frac{c^2 \log(f \vee m)}{C_1 + C_2 c \sqrt{\log(f \vee m)/n}} \right) \\ &\leq (f^2 + f) \exp \left( - \frac{c^2 \log(f \vee m)}{C_1 + C_2/\kappa} \right) \leq \frac{f^2 + f}{\max(m, f)^4} \end{aligned}$$

where in the last two inequalities, we used the fact that  $c^2 \geq 4(C_1 + C_2/\kappa)$  and  $n \geq (\kappa c)^2 \log \max(m, f)$ , and applied the Bernstein's inequality. Similarly, we have for  $\varepsilon = cD' \sqrt{\frac{\log(f \vee m)}{n}}$

$$\begin{aligned} &\mathbb{P} \left( \left| \text{diag}^{-1/2}(B_0) \frac{1}{\sqrt{m}} \sum_{k=1}^m \sum_{\ell=1}^m \left( \hat{S}_n^{k\ell} M_{k\ell} - a_{k\ell} M_{k\ell} B_0 \right) \text{diag}^{-1/2}(B_0) \right|_{\max} \geq \varepsilon' \right) \\ &\leq \frac{m^2 + m}{\max(m, f)^4}. \end{aligned}$$

The theorem is thus proved for  $n \geq (\kappa c)^2 \log \max(m, f) \geq 4(\kappa^2 C_1 + C_2 \kappa) \log \max(m, f)$  where  $\kappa > 0$  can be chosen to be small enough so that the lower bound reaches  $\log \max(m, f)$ .  $\square$

## D Poof of Theorem 4.3 and Corollary 4.4

Let  $\Theta_0 := \rho(A_0)^{-1} \succ 0$  and  $\Phi_0 := \rho(B_0)^{-1} \succ 0$ . Let  $\Theta_0 = (\theta_{ij})$  and  $\Phi_0 = (\phi_{ij})$ . Then,

$$\begin{aligned} 0 < \varphi_{\min}(\rho(A_0)) \leq 1 \leq \varphi_{\max}(\rho(A_0)) < +\infty \text{ and hence } \kappa(\rho(A_0)) \geq 1, \\ \text{and } 0 < \varphi_{\min}(\rho(B_0)) \leq 1 \leq \varphi_{\max}(\rho(B_0)) < +\infty \text{ and hence } \kappa(\rho(B_0)) \geq 1, \end{aligned}$$

given that  $\sum_{i=1}^m \varphi_i(\rho(A_0)) = \text{tr}(\rho(A_0)) = m$  and  $\sum_{i=1}^f \varphi_i(\rho(B_0)) = f$ .

We first write our estimator  $\hat{\Theta}$  for  $\rho(A_0)^{-1}$  and  $\hat{\Phi}$  for  $\rho(B_0)^{-1}$  as follows:

$$\hat{\Theta} = \arg \min_{\Theta \succ 0} \{ \text{tr}(\Theta \hat{\Gamma}(A_0)) - \log |\Theta| + \lambda_B |\Theta|_{1,\text{off}} \}, \quad (52a)$$

$$\hat{\Phi} = \arg \min_{\Phi \succ 0} \{ \text{tr}(\Phi \hat{\Gamma}(B_0)) - \log |\Phi| + \lambda_A |\Phi|_{1,\text{off}} \}, \quad (52b)$$

where  $\lambda_A, \lambda_B$  are non-negative regularization parameters, and  $\hat{\Gamma}(A_0)$  and  $\hat{\Gamma}(B_0)$  are sample correlation matrices. A unique minimizer exists for each objective function above; see [Ravikumar et al. \(2008\)](#). Then  $\hat{A} := \hat{\Theta}^{-1}$  and  $\hat{B} := \hat{\Phi}^{-1}$ , where  $\hat{A}$  and  $\hat{B}$  are the unique minimizers as defined in Theorem 4.3. Let us define the following shorthand notation:

$$S_{A_0} = \{(i, j) : \theta_{ij} \neq 0, i \neq j\} \quad \text{and} \quad S_{B_0} = \{(i, j) : \phi_{ij} \neq 0, i \neq j\}.$$

Let  $\Delta_{A_0} := \hat{\Theta} - \Theta_0$  and  $\Delta_{B_0} := \hat{\Phi} - \Phi_0$ .

We need the following lemma. For an index set  $S$  and a matrix  $W = [w_{ij}]$ , write  $W_S \equiv (w_{ij} I((i, j) \in S))$ , where  $I(\cdot)$  is an indicator function.

**Lemma D.1.** *Let  $\Theta_0 \succ 0$ . Let  $S = \{(i, j) : \Theta_{0ij} \neq 0, i \neq j\}$  and  $S^c = \{(i, j) : \Theta_{0ij} = 0, i \neq j\}$ . Then for all  $\Delta \in \mathbf{R}^{m \times m}$ , we have*

$$|\Theta_0 + \Delta|_{1,\text{off}} - |\Theta_0|_{1,\text{off}} \geq |\Delta_{S^c}|_1 - |\Delta_S|_1 \quad (53)$$

Moreover, we have on event  $\mathcal{T}(A_0)$ , for all  $\Delta \in \mathbf{R}^{m \times m}$ ,

$$\left| \text{tr}(\Delta(\hat{\Gamma}(A_0) - \rho(A_0))) \right| \leq \delta_{n,f} |\Delta|_{1,\text{off}} = \delta_{n,f} (|\Delta_{S^c}|_1 + |\Delta_S|_1). \quad (54)$$

*Proof.* We write  $\Theta_0 = \text{diag}(\Theta_0) + \Theta_{0,S} + \Theta_{0,S^c}$  and observe that  $\Theta_{0,S^c} = \underline{0}$  and hence  $|\Theta_0|_{1,\text{off}} = |\Theta_{0,S}|_1 = |\Theta_{0,S}|_{1,\text{off}}$ . Thus

$$\begin{aligned} |\Theta_0 + \Delta|_{1,\text{off}} &= |\Theta_{0,S} + \Delta_S|_{1,\text{off}} + |\Delta_{S^c}|_{1,\text{off}}, \quad \text{hence} \\ |\Theta_0 + \Delta|_{1,\text{off}} - |\Theta_0|_{1,\text{off}} &\geq |\Theta_{0,S} + \Delta_S|_{1,\text{off}} - |\Theta_{0,S}|_{1,\text{off}} + |\Delta_{S^c}|_{1,\text{off}} \end{aligned} \quad (55)$$

$$\geq |\Delta_{S^c}|_{1,\text{off}} - |\Delta_S|_{1,\text{off}} = |\Delta_{S^c}|_1 - |\Delta_S|_1 \quad (56)$$

where (55) follows from the triangle inequality and (56) follows from definition of  $S$  and  $S^c$ . The “moreover” part follows from Lemma 4.5 and definition of event  $\mathcal{T}(A_0)$ .  $\square$

Proposition D.2 is a standard result; see [Zhou et al. \(2008\)](#) for its proof.

**Proposition D.2.** *Let  $B$  be a  $p \times p$  matrix. If  $B \succ 0$  and  $B + D \succ 0$ , then  $B + vD \succ 0$  for all  $v \in [0, 1]$ .*



## D.1 Proof of Theorem 4.3

The proof follows arguments in Rothman et al. (2008), and hence omitted.  $\square$

*Proof of Corollary 4.4.* Let  $\widehat{\Gamma}(A_0)$  be the sample correlation matrix. Let  $\lambda := \lambda_B$ . Let  $\widehat{\Theta}$  be the optimal solution to (52a). Let  $\Theta_0 := \rho(A_0)^{-1}$ . Then  $\Delta = \widehat{\Theta} - \Theta_0$  minimizes  $G(\Delta)$  defined as follows:

$$\begin{aligned} G(\Delta) &:= \log |\Theta_0| - \log |\Theta_0 + \Delta| + \text{tr}(\Delta \rho(A_0)) + \text{tr}(\Delta(\widehat{\Gamma}(A_0) - \rho(A_0))) \\ &\quad + \lambda(|\Theta_0 + \Delta|_{1,\text{off}} - |\Theta_0|_{1,\text{off}}) \\ &= A + \text{tr}(\Delta(\widehat{\Gamma}(A) - \rho(A_0))) + \lambda_{n,f}(|\Theta_0 + \Delta|_{1,\text{off}} - |\Theta_0|_{1,\text{off}}) \end{aligned} \quad (57)$$

$$\geq A - \delta_{n,f} |\Delta_{S^c}|_1 - \delta_{n,f} |\Delta_S|_1 + \lambda_{n,f} |\Delta_{S^c}|_1 - \lambda_{n,f} |\Delta_S|_1 \quad (58)$$

$$= A - (\delta_{n,f} + \lambda_{n,f}) |\Delta_S|_1 + (\lambda_{n,f} - \delta_{n,f}) |\Delta_{S^c}|_1 \quad (59)$$

where  $A = \text{vec}\{\Delta\}^T \left( \int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv \right) \text{vec}\{\Delta\}$ , and in (58), we have applied (53) and (54).

Clearly  $G(0) = 0$  and hence  $G(\Delta) \leq G(0) = 0$ . Hence for  $\lambda_{n,f} := \lambda_B \geq \delta_{n,f}/\varepsilon$ , we have by (59)

$$\begin{aligned} -A + (\delta_{n,f} + \lambda_{n,f}) |\Delta_S|_1 &\geq (\lambda_{n,f} - \delta_{n,f}) |\Delta_{S^c}|_1 \geq (1 - \varepsilon) \lambda_{n,f} |\Delta_{S^c}|_1 \\ \text{and hence } (1 - \varepsilon) \lambda_{n,f} |\Delta_{S^c}|_1 &\leq -A + (\delta_{n,f} + \lambda_{n,f}) |\Delta_S|_1 \\ &\leq -A + (1 + \varepsilon) \lambda_{n,f} |\Delta_S|_1 \end{aligned}$$

Thus we have  $|\Delta_{S^c}|_1 \leq \frac{1+\varepsilon}{1-\varepsilon} |\Delta_S|_1$  given that  $A \geq 0$ ; To see this, we note that  $\Theta_0 \succ 0$ , and hence  $\Theta_0 + v\Delta \succ 0$  for all  $v \in [0, 1]$  in view of Proposition D.2, given that  $\widehat{\Theta} = \Theta_0 + \Delta \succ 0$  as an optimal solution to (52a). To see the last inequality, we have

$$\begin{aligned} \left| \widehat{\Theta} - \Theta_0 \right|_{1,\text{off}} = |\Delta_{A_0}|_{1,\text{off}} &\leq \frac{2}{1-\varepsilon} |\Delta_S|_1 \leq \frac{2}{1-\varepsilon} \sqrt{|A_0^{-1}|_{0,\text{off}}} \|\Delta_S\|_F \\ &\leq \frac{2}{1-\varepsilon} \sqrt{|A_0^{-1}|_{0,\text{off}}} \|\Delta_{A_0}\|_F, \end{aligned}$$

and it holds by plugging in the bound on  $\|\Delta_{A_0}\|_F$ .  $\square$

## D.2 Discussion

To get a sense of how tightly we can bound these entries in  $\Delta_{A_0}$ , we do the following calculations. For  $\varepsilon = 2/3$ , we have  $|\Delta_{S^c}|_1 \leq 5 |\Delta_S|_1$ , and hence

$$\begin{aligned} \left| \widehat{\Theta} - \Theta_0 \right|_{1,\text{off}} = |\Delta_{A_0}|_{1,\text{off}} &\leq 6 |\Delta_S|_1 \leq 6 \sqrt{|A_0^{-1}|_{0,\text{off}}} \|\Delta_S\|_F \\ &\leq 6 \sqrt{|A_0^{-1}|_{0,\text{off}}} \frac{9(1+\varepsilon)\lambda_B \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1}{2\varphi_{\min}^2(\rho(A_0))} \\ &\leq 27 \frac{1+\varepsilon}{\varphi_{\min}^2(\rho(A_0))} \lambda_B \left( |A_0^{-1}|_{0,\text{off}} \vee 1 \right). \end{aligned}$$

From Corollary 4.4, we have for  $\Theta_0 = \rho(A_0)^{-1}$  and for  $S_{A_0} \neq \emptyset$ ,

$$\frac{|\Delta_{A_0}|_{1,\text{off}}}{|A_0^{-1}|_{0,\text{off}}} \leq 27 \frac{1+\varepsilon}{\varphi_{\min}^2(\rho(A_0))} \lambda_B \asymp \frac{\delta_{n,f}}{\varphi_{\min}^2(\rho(A_0))} \rightarrow 0.$$

Such results are useful for bounding the number of falsely selected edges and the number of falsely deleted edges under certain separation assumption, which states that: the minimum absolute value of  $\theta_{ij}$ :  $\min_{i,j,i \neq j} |\theta_{ij}|$  among all non-zero off-diagonal entries in  $\Theta_0$  is bounded away from 0. We do not pursue such refinements in this work. On the other hand, for the diagonal part of  $\Delta_{A_0}$ , we have

$$\begin{aligned} |\text{diag}(\Delta_{A_0})|_1 &\leq \sqrt{m} \|\Delta_{A_0}\|_F \leq \frac{9}{2} \frac{1+\varepsilon}{\varphi_{\min}^2(\rho(A_0))} \lambda_B \sqrt{m(|A_0^{-1}|_{0,\text{off}} \vee 1)} \\ \text{and hence } \frac{|\text{diag}(\Delta_{A_0})|_1}{m} &\leq \frac{9}{2} \frac{1+\varepsilon}{\varphi_{\min}^2(\rho(A_0))} \lambda_B \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee 1}{m}} \\ &\asymp \frac{\delta_{n,f}}{\varphi_{\min}^2(\rho(A_0))} \sqrt{\frac{|A_0^{-1}|_{0,\text{off}} \vee 1}{m}} \rightarrow 0 \end{aligned}$$

at a faster (or comparable) rate in terms of  $\ell_1$  norm in average than that for the off-diagonal entries when  $|S_{A_0}| < m$  (or when  $|S_{A_0}| \asymp m$ ).

## E An outline for convergence in the operator and the Frobenius norm

We need some auxiliary results for proving Theorems A.1 and A.2. Throughout this section, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. It is clear from the proof of Theorem 4.1 that  $\mathcal{T}(A_0) \cap \mathcal{T}(B_0)$  holds on event  $\mathcal{X}_0$ , and thus all statements in Theorem 4.3 hold on  $\mathcal{X}_0$ . Let  $\widehat{W}_1$  and  $\widehat{W}_2$  be as defined in (7a) and (7b). Claim E.1 provides the large deviation bounds on the spectral norm for estimating  $W_1$  and  $W_2$  using  $\widehat{W}_1$  and  $\widehat{W}_2$ .

**Claim E.1.** Denote by  $\beta_n := \frac{f\tau'_0}{\text{tr}(B_0)}$  and  $\alpha_n := \frac{m\tau_0}{\text{tr}(A_0)}$ , where  $\tau_0$  and  $\tau'_0$  are as defined in Theorem 4.1. Let  $1 > \lambda_{A_0} > 3\alpha_n$  and  $1 > \lambda_{B_0} > 3\beta_n$ . Let  $W_1^2/\text{tr}(B_0) = \text{diag}(a_{11}, \dots, a_{mm})$  and  $W_2^2/\text{tr}(A_0) = \text{diag}(b_{11}, \dots, b_{ff})$ . Then on  $\mathcal{X}_0$ , we have for  $\beta'_n := \frac{\beta_n}{\sqrt{1-\beta_n}} \leq \frac{\lambda_{B_0}}{\sqrt{6}}$  and  $\alpha'_n = \frac{\alpha_n}{\sqrt{1-\alpha_n}} \leq \frac{\lambda_{A_0}}{\sqrt{6}}$ ,

$$\begin{aligned} \|\widehat{W}_1 - W_1\|_2 &\leq \beta_n \sqrt{\text{tr}(B_0)} \sqrt{a_{\max}}, \\ \|\widehat{W}_1^{-1} - W_1^{-1}\|_2 &\leq \beta'_n / \sqrt{\text{tr}(B_0)} \sqrt{a_{\min}}, \\ \|\widehat{W}_2 - W_2\|_2 &\leq \alpha_n \sqrt{\text{tr}(A_0)} \sqrt{b_{\max}}, \\ \text{and } \|\widehat{W}_2^{-1} - W_2^{-1}\|_2 &\leq \alpha'_n / \sqrt{\text{tr}(A_0)} \sqrt{b_{\min}}. \end{aligned}$$

We next state the following convenient results.

**Lemma E.2.** Let  $\|\cdot\|$  be a matrix norm which satisfies the triangle inequality and is multiplicative with respect to the Kronecker product:  $\|A \otimes B\| = \|A\| \|B\|$ . Then for  $\Delta_A := A_1 - A$  and  $\Delta_B := B_1 - B$ ,

$$\|A_1 \otimes B_1 - A \otimes B\| \leq \|\Delta_A\| \|B\| + \|\Delta_B\| \|A\| + \|\Delta_A\| \|\Delta_B\|.$$

It is clear from Lemma E.2 that the intermediate results in Lemma E.3 are useful in bounding the operator norm and the Frobenius norm on  $\Delta$  and  $\Delta'$  to be defined in (60) and (61).

**Lemma E.3.** Suppose (A1) and (A2) hold. Let  $\alpha_n < \lambda_{A_0}/3$  and  $\beta_n < \lambda_{B_0}/3$  where  $\lambda_{A_0}, \lambda_{B_0} < 1$ . We have on  $\mathcal{X}_0$  for some absolute constants  $9 < C, C' < 18$ ,

$$\begin{aligned} \left\| \widehat{W}_1 \widehat{A} \widehat{W}_1 / \text{tr}(B_0) - A_0 \right\|_2 &\leq 2C \lambda_{B_0} a_{\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}, \\ \text{and } \left\| \widehat{W}_2 \widehat{B} \widehat{W}_2 / \text{tr}(A_0) - B_0 \right\|_2 &\leq 2C' \lambda_{A_0} b_{\max} \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}; \end{aligned}$$

and for some  $5 < C, C' < 19/2$ ,

$$\begin{aligned} \left\| \text{tr}(B_0) \left( \widehat{W}_1 \widehat{A} \widehat{W}_1 \right)^{-1} - A_0^{-1} \right\|_2 &\leq \frac{2C \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} \varphi_{\min}^2(\rho(A_0))}, \\ \text{and } \left\| \text{tr}(A_0) \left( \widehat{W}_2 \widehat{B} \widehat{W}_2 \right)^{-1} - B_0^{-1} \right\|_2 &\leq \frac{2C' \lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B_0))}. \end{aligned}$$

On  $\mathcal{X}_0$ , for some absolute constants  $18 > C, C' > 9$ ,

$$\begin{aligned} \left\| \widehat{W}_1 \widehat{A} \widehat{W}_1 / \text{tr}(B_0) - A_0 \right\|_F &\leq 2C \lambda_{B_0} a_{\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}, \\ \text{and } \left\| \widehat{W}_2 \widehat{B} \widehat{W}_2 / \text{tr}(A_0) - B_0 \right\|_F &\leq 2C' \lambda_{A_0} b_{\max} \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}; \end{aligned}$$

and for some  $19/2 > C, C' > 5$ ,

$$\begin{aligned} \left\| \text{tr}(B_0) \left( \widehat{W}_1 \widehat{A} \widehat{W}_1 \right)^{-1} - A_0^{-1} \right\|_F &\leq \frac{2C \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}}{a_{\min} \varphi_{\min}^2(\rho(A_0))}, \\ \text{and } \left\| \text{tr}(A_0) \left( \widehat{W}_2 \widehat{B} \widehat{W}_2 \right)^{-1} - B_0^{-1} \right\|_F &\leq \frac{2C' \lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}}{b_{\min} \varphi_{\min}^2(\rho(B_0))}. \end{aligned}$$

Let us denote by  $\Delta$  and  $\Delta'$  the following:

$$\Delta := \widehat{W}_1 \widehat{A} \widehat{W}_1 \otimes \widehat{W}_2 \widehat{B} \widehat{W}_2 / (\text{tr}(B_0) \text{tr}(A_0)) - A_0 \otimes B_0, \quad (60)$$

$$\Delta' := \text{tr}(B_0) \text{tr}(A_0) \left( \widehat{W}_1 \widehat{A} \widehat{W}_1 \right)^{-1} \otimes \left( \widehat{W}_2 \widehat{B} \widehat{W}_2 \right)^{-1} - A_0^{-1} \otimes B_0^{-1}. \quad (61)$$

**Lemma E.4.** Let  $\|\cdot\|$  be a matrix norm which satisfies the triangle inequality and is multiplicative with respect to the Kronecker product:  $\|A \otimes B\| = \|A\| \|B\|$ . Then for  $\widehat{A} \otimes \widehat{B}$  as defined in (8), we have for  $\Sigma_0 = A_0 \otimes B_0$ ,

$$\left\| \widehat{A \otimes B}^{-1} - \Sigma_0^{-1} \right\| \leq (\alpha_n \wedge \beta_n) \|A_0^{-1}\| \|B_0^{-1}\| + (1 + \alpha_n \wedge \beta_n) \|\Delta'\|, \quad (62)$$

$$\left\| \widehat{A \otimes B} - \Sigma_0 \right\| \leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|A_0\| \|B_0\| + \left(1 + \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2}\right) \|\Delta\|. \quad (63)$$

## F Proofs for the rates of convergence for the Gemini estimators

### F.1 Proof of Theorem A.1

Throughout this proof, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. First we bound  $\|\Delta'\|_2$  using Lemmas E.2 and E.3 as follows. Let  $C, C'$  be some absolute constants such that  $19/2 > C, C' > 5$ , and  $\widehat{W}_1$  and  $\widehat{W}_2$  be as defined in (7a) and (7b) respectively.

Then on  $\mathcal{X}_0$ ,

$$\begin{aligned} \|\Delta'\|_2 &\leq \left\| \left( \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} \right)^{-1} - A_0^{-1} \right\|_2 \|B_0^{-1}\|_2 + \|A_0^{-1}\|_2 \left\| \left( \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} \right)^{-1} - B_0^{-1} \right\|_2 \\ &\quad + \left\| \text{tr}(B_0) \left( \widehat{W}_1^{-1} \widehat{A}^{-1} \widehat{W}_1^{-1} \right) - A_0^{-1} \right\|_2 \left\| \text{tr}(A_0) \left( \widehat{W}_2^{-1} \widehat{B}^{-1} \widehat{W}_2^{-1} \right) - B_0^{-1} \right\|_2 \\ &\leq \frac{2C\lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{a_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}(B_0)} + \frac{2C'\lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{b_{\min} \varphi_{\min}^2(\rho(B_0)) \varphi_{\min}(A_0)} + \\ &\quad + \frac{4CC'\lambda_{A_0} \lambda_{B_0}}{a_{\min} b_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}^2(\rho(B_0))} \cdot \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}, \end{aligned} \quad (64)$$

and for  $\|\Delta'\|_2$  bounded as in (64), we have by (52), (52), (42a), and (42b)

$$\begin{aligned} (\alpha_n \wedge \beta_n) \|\Delta'\|_2 &\leq \frac{4CC'\lambda_{A_0} \lambda_{B_0} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{3a_{\min} b_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}^2(\rho(B_0))} \times \\ &\quad \left( \frac{\varphi_{\min}(\rho(B_0))}{2C' \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}} + \frac{\varphi_{\min}(\rho(A_0))}{2C \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}} + \lambda_{A_0} \wedge \lambda_{B_0} \right) \\ &\leq \frac{8CC'\lambda_{A_0} \lambda_{B_0}}{5a_{\min} b_{\min}} \left( \frac{\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}^2(\rho(A_0))} \frac{\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}^2(\rho(B_0))} \right). \end{aligned} \quad (65)$$

The bound in the theorem statement for estimating  $A_0^{-1} \otimes B_0^{-1}$  thus holds by inserting (64) and (65) in (62). Next we bound the error in the operator norm for estimating  $A_0 \otimes B_0$ . Let  $9 < C, C' < 18$ . On  $\mathcal{X}_0$ , we have by Lemma E.2 and Lemma E.3,

$$\begin{aligned} \|\Delta\|_2 &= \left\| \left( \frac{\widehat{W}_1}{\sqrt{\text{tr}(B_0)}} \right) \widehat{A} \left( \frac{\widehat{W}_1}{\sqrt{\text{tr}(B_0)}} \right) \otimes \left( \frac{\widehat{W}_2}{\sqrt{\text{tr}(A_0)}} \right) \widehat{B} \left( \frac{\widehat{W}_2}{\sqrt{\text{tr}(A_0)}} \right) - A_0 \otimes B_0 \right\|_2 \\ &\leq \left\| \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} - A_0 \right\|_2 \|B_0\|_2 + \|A_0\|_2 \left\| \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} - B_0 \right\|_2 + \left\| \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} - A_0 \right\|_2 \left\| \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} - B_0 \right\|_2 \\ &\leq 2C\lambda_{B_0} a_{\max} \|B_0\|_2 \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} + 2C'\lambda_{A_0} b_{\max} \|A_0\|_2 \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} \\ &\quad + 4CC'\lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}, \end{aligned}$$

and hence by (52), (52), and (42e),

$$\begin{aligned} \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|\Delta\|_2 &\leq 2CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 + \\ &\quad \left( \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1 \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1 \right) \times \left( \frac{1}{2C} + \frac{1}{2C'} + (\lambda_{A_0} \wedge \lambda_{B_0}) \right) \\ &\leq \frac{20CC'}{9} \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee 1 \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee 1. \end{aligned}$$

The theorem is thus proved by inserting the bounds immediately above in (63).  $\square$

## F.2 Proof of Theorem A.2

Throughout this proof, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. We first obtain the rate of convergence in the Frobenius norm for estimating  $A_0 \otimes B_0$ . First we bound  $\|\Delta\|_F$  using Lemma E.2 and Lemma E.3 as follows: for  $9 < C, C' < 18$ ,

$$\begin{aligned} \|\Delta\|_F &\leq \|B_0\|_F \left\| \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} - A_0 \right\|_F + \|A_0\|_F \left\| \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} - B_0 \right\|_F + \left\| \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} - A_0 \right\|_F \left\| \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} - B_0 \right\|_F \\ &\leq 2C \lambda_{B_0} a_{\max} \|B_0\|_F \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m + 2C' \lambda_{A_0} b_{\max} \|A_0\|_F \kappa(\rho(B_0))^2 \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f \\ &\quad + 4CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f. \end{aligned}$$

Hence for  $\|B_0\|_F \leq \sqrt{f} \|B_0\|_2$ , and  $\|A_0\|_F \leq \sqrt{m} \|A_0\|_2$ , we have by (42e), (52), and (52),

$$\begin{aligned} \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|\Delta\|_F &\leq 2CC' \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \cdot \\ &\quad \left( \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f \right) \times \left( \frac{1}{2C} + \frac{1}{2C'} + (\lambda_{A_0} \wedge \lambda_{B_0}) \right) \\ &\leq \frac{20CC'}{9} \lambda_{A_0} \lambda_{B_0} a_{\max} b_{\max} \kappa(\rho(A_0))^2 \kappa(\rho(B_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f. \end{aligned}$$

The bound on  $\left\| \widehat{A} \widehat{B} - A_0 \otimes B_0 \right\|_F$  thus holds by inserting the bounds above in (63). We next obtain the rate of convergence in the Frobenius norm for estimating  $A_0^{-1} \otimes B_0^{-1}$ . Let  $19/2 > C, C' > 5$ . We bound  $\|\Delta'\|_F$  using Lemma E.2 and E.3 as follows,

$$\begin{aligned} \|\Delta'\|_F &\leq \|B_0^{-1}\|_F \left\| \left( \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} \right)^{-1} - A_0^{-1} \right\|_F + \|A_0^{-1}\|_F \left\| \left( \frac{\widehat{W}_2 \widehat{B} \widehat{W}_2}{\text{tr}(A_0)} \right)^{-1} - B_0^{-1} \right\|_F \\ &\quad + \left\| \left( \widehat{W}_1 \widehat{A} \widehat{W}_1 / \text{tr}(B_0) \right)^{-1} - A_0^{-1} \right\|_F \left\| \left( \widehat{W}_2 \widehat{B} \widehat{W}_2 / \text{tr}(A_0) \right)^{-1} - B_0^{-1} \right\|_F \\ &\leq \frac{2C \lambda_{B_0} \|B_0^{-1}\|_F \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m}{a_{\min} \varphi_{\min}^2(\rho(A_0))} + \frac{2C' \lambda_{A_0} \|A_0^{-1}\|_F \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f}{b_{\min} \varphi_{\min}^2(\rho(B_0))} + \\ &\quad + \frac{4CC' \lambda_{A_0} \lambda_{B_0}}{a_{\min} b_{\min} \varphi_{\min}^2(\rho(A_0)) \varphi_{\min}^2(\rho(B_0))} \cdot \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f. \end{aligned}$$

And thus by (42a), (42b), (52), and (52)

$$\begin{aligned}
\frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} \|\Delta'\|_F &\leq \frac{2C\lambda_{B_0}\lambda_{A_0}\sqrt{|A_0^{-1}|_{0,\text{off}} \vee m\sqrt{f}}}{3a_{\min}b_{\min}\varphi_{\min}^2(\rho(A_0))\varphi_{\min}(\rho(B_0))} \\
&+ \frac{2C'\lambda_{A_0}\lambda_{B_0}\sqrt{|B_0^{-1}|_{0,\text{off}} \vee f\sqrt{m}}}{3a_{\min}b_{\min}\varphi_{\min}^2(\rho(B_0))\varphi_{\min}(\rho(A_0))} + \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3} . \\
&\frac{4CC'\lambda_{A_0}\lambda_{B_0}}{a_{\min}b_{\min}\varphi_{\min}^2(\rho(A_0))\varphi_{\min}^2(\rho(B_0))} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f} \\
&\leq \frac{8CC'\lambda_{A_0}\lambda_{B_0}}{5a_{\min}b_{\min}\varphi_{\min}^2(\rho(A_0))\varphi_{\min}^2(\rho(B_0))} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee f}.
\end{aligned}$$

The theorem is thus proved by putting things together in (62).  $\square$

### F.3 Proof of Corollary A.3

Throughout this proof, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. We first note that the bounds on  $\|\widehat{B}_* - B_*\|$  and  $\|\widehat{B}_*^{-1} - B_*^{-1}\|$  follow from Lemma E.3 immediately: for  $W_2 = \sqrt{\text{tr}(A_0)}\text{diag}(B_0)^{1/2}$  and  $\widehat{W}_2$  as defined in (7b)

$$\|\widehat{B}_* - B_*\| = \frac{1}{m} \|\widehat{W}_2 \widehat{B} \widehat{W}_2 - \text{tr}(A_0)B_0\| = \frac{\text{tr}(A_0)}{m} \|\widehat{W}_2 \widehat{B} \widehat{W}_2 / \text{tr}(A_0) - B_0\|.$$

We can now plug in the bounds on the operator and the Frobenius norm from Lemma E.3. The proof for  $\widehat{A}_*$  is long and monotonous, and hence omitted.  $\square$

### F.4 Proof of Claim E.1

We note that the following bounds  $\forall i = 1, \dots, m$  hold on  $\mathcal{X}_0$ , for  $\beta := \beta_n$ ,

$$\begin{aligned}
-\frac{f\tau'_0}{\text{tr}(B_0)} &\leq \left( \sqrt{\frac{\frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2}{a_{ii}\text{tr}(B_0)}} \right)^2 - 1 \leq \frac{f\tau'_0}{\text{tr}(B_0)} := \beta \quad \text{and} \\
\max_{i=1,\dots,m} \left| \frac{\sqrt{\frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2}}{\sqrt{a_{ii}\text{tr}(B_0)}} - 1 \right| &\leq (1 - \sqrt{1 - \beta}) \vee (\sqrt{1 + \beta} - 1) \leq \beta.
\end{aligned}$$

Thus the first inequality holds. Similarly, on  $\mathcal{X}_0$ , we obtain

$$\begin{aligned}
\forall i = 1, \dots, m, \quad \frac{1}{\sqrt{1 + \beta}} &\leq \frac{\sqrt{a_{ii}\text{tr}(B_0)}}{\sqrt{\frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2}} \leq \frac{1}{\sqrt{1 - \beta}} \quad \text{and hence} \\
\left| \frac{\sqrt{a_{ii}\text{tr}(B_0)}}{\sqrt{\frac{1}{n} \sum_{t=1}^n \|x(t)^i\|_2^2}} - 1 \right| &\leq \left( \frac{\sqrt{1 + \beta} - 1}{\sqrt{1 + \beta}} \right) \vee \left( \frac{1 - \sqrt{1 - \beta}}{\sqrt{1 - \beta}} \right) \leq \frac{\beta}{\sqrt{1 - \beta}}.
\end{aligned}$$

Thus the second inequality also holds. The statements about  $\widehat{W}_2$  and  $\widehat{W}_2^{-1}$  also hold following the same line of arguments.  $\square$

## F.5 Proof of Lemma E.2

We have by distributivity of the Kronecker product,

$$\begin{aligned} A_1 \otimes B_1 &= (A + \Delta_A) \otimes (B + \Delta_B) = A \otimes (B + \Delta_B) + \Delta_A \otimes (B + \Delta_B) \\ &= A \otimes B + A \otimes \Delta_B + \Delta_A \otimes B + \Delta_A \otimes \Delta_B, \end{aligned}$$

hence by the triangle inequality, we have

$$\begin{aligned} \|A_1 \otimes B_1 - A \otimes B\| &= \|A \otimes \Delta_B + \Delta_A \otimes B + \Delta_A \otimes \Delta_B\| \\ &\leq \|A \otimes \Delta_B\| + \|\Delta_A \otimes B\| + \|\Delta_A \otimes \Delta_B\| \\ &= \|\Delta_B\| \|A\| + \|\Delta_A\| \|B\| + \|\Delta_A\| \|\Delta_B\|, \end{aligned}$$

where the last step follows from the multiplicativity of the norm with respect to the Kronecker product.  $\square$

## F.6 Proof of Lemma E.3

In the following proofs, we will only derive the bounds for estimating  $A_0$  and  $A_0^{-1}$ , as the bounds for  $B_0$  and  $B_0^{-1}$  are derived in exactly the same manner, and hence omitted. We need the following proposition.

**Proposition F.1.** *Let  $\widehat{W}$  and  $W$  be diagonal positive definite matrices. Let  $\widehat{\Psi}$  and  $\Psi$  be symmetric positive definite matrices. Then*

$$\begin{aligned} \|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\|_2 &\leq \left(\|\widehat{W} - W\|_2 + \|W\|_2\right)^2 \|\widehat{\Psi} - \Psi\|_2 \\ &\quad + \|\widehat{W} - W\|_2 \left(\|\widehat{W} - W\|_2 + 2\right) \|\Psi\|_2 \\ \|\widehat{W}\widehat{\Psi}\widehat{W} - W\Psi W\|_F &\leq \left(\|\widehat{W} - W\|_2 + \|W\|_2\right)^2 \|\widehat{\Psi} - \Psi\|_F \\ &\quad + \|\widehat{W} - W\|_2 \left(\|\widehat{W} - W\|_2 + 2\right) \|\Psi\|_F. \end{aligned}$$

*Proof of Lemma E.3.* Throughout this proof, we assume that event  $\mathcal{X}_0$  as defined in Lemma 4.2 holds. Let diagonal matrix  $W := W_1/\sqrt{\text{tr}(B_0)}$  and  $\widehat{W} := \widehat{W}_1/\sqrt{\text{tr}(B_0)}$ . By Proposition F.1, Claim E.1, and Theorem 4.3, we have for  $\beta := \beta_n \leq \lambda_{B_0}/3 < 1/3$  and  $18 > C > 9$ ,

$$\begin{aligned} \|\widehat{W}_1 \widehat{A} \widehat{W}_1 / \text{tr}(B_0) - A_0\|_2 &= \|\widehat{W} \widehat{A} \widehat{W} - \text{diag}(A_0)^{1/2} \rho(A_0) \text{diag}(A_0)^{1/2}\|_2 \\ &\leq (1 + \beta)^2 a_{\max} \|\widehat{A} - \rho(A_0)\|_2 + (\beta^2 + 2\beta) a_{\max} \|\rho(A_0)\|_2 \\ &\leq 2C \lambda_{B_0} a_{\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0, \text{off}}} \vee 1, \end{aligned}$$



where we used the fact that  $\kappa(\rho(A_0)) \geq \|\rho(A_0)\|_2 \geq 1$ , given that  $\varphi_{\min}(\rho(A_0)) \leq 1$  as shown in (52); and

$$\begin{aligned} \left\| \widehat{W}_1 \widehat{A} \widehat{W}_1 / \text{tr}(B_0) - A_0 \right\|_F &\leq (1 + \beta)^2 a_{\max} \left\| \widehat{A} - \rho(A_0) \right\|_F + (\beta^2 + 2\beta) a_{\max} \|\rho(A_0)\|_F \\ &\leq 2C \lambda_{B_0} a_{\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0, \text{off}}} \vee m. \end{aligned}$$

Similarly, for  $\beta' := \beta_n / \sqrt{1 - \beta_n} \leq \lambda_{B_0} / \sqrt{6}$ , where  $\beta_n < 1/3$ , we have by Proposition F.1, Claim E.1, and Theorem 4.3,

$$\begin{aligned} \left\| \left( \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} \right)^{-1} - A_0^{-1} \right\|_2 &\leq \frac{(1 + \beta')^2}{a_{\min}} \left\| \widehat{A}^{-1} - \rho(A_0)^{-1} \right\|_2 + \frac{(\beta' + 2)\beta'}{a_{\min}} \|\rho(A_0)^{-1}\|_2 \\ &\leq (2C + 1) \lambda_{B_0} \sqrt{|A_0^{-1}|_{0, \text{off}}} \vee 1 / (a_{\min} \varphi_{\min}^2(\rho(A_0))), \\ \left\| \left( \frac{\widehat{W}_1 \widehat{A} \widehat{W}_1}{\text{tr}(B_0)} \right)^{-1} - A_0^{-1} \right\|_F &\leq \frac{(1 + \beta')^2}{a_{\min}} \left\| \widehat{A}^{-1} - \rho(A_0)^{-1} \right\|_F + \frac{(\beta' + 2)\beta'}{a_{\min}} \|\rho(A_0)^{-1}\|_F \\ &\leq (2C + 1) \lambda_{B_0} \sqrt{|A_0^{-1}|_{0, \text{off}}} \vee m / (a_{\min} \varphi_{\min}^2(\rho(A_0))), \end{aligned}$$

where  $9/2 < C < 9$  and hence the statement in the Lemma holds for  $5 < C < 19/2$ .  $\square$

*Proof of Proposition F.1.* By the triangle inequality, we have

$$\begin{aligned} \left\| \widehat{W} \widehat{\Psi} \widehat{W} - W \Psi W \right\|_2 &= \\ &\left\| (\widehat{W} - W) \widehat{\Psi} (\widehat{W} - W) + W \widehat{\Psi} (\widehat{W} - W) + (\widehat{W} - W) \widehat{\Psi} W + W (\widehat{\Psi} - \Psi) W \right\|_2 \\ &\leq \left( \left\| \widehat{W} - W \right\|_2^2 + 2 \left\| \widehat{W} - W \right\|_2 \|W\|_2 \right) \|\Psi\|_2 + \left( \left\| \widehat{W} - W \right\|_2 + \|W\|_2 \right)^2 \left\| \widehat{\Psi} - \Psi \right\|_2. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \left\| \widehat{W} \widehat{\Psi} \widehat{W} - W \Psi W \right\|_F &\leq \\ &\left\| \widehat{W} - W \right\|_2^2 \left\| \widehat{\Psi} \right\|_F + 2 \left\| \widehat{W} - W \right\|_2 \|W\|_2 \left\| \widehat{\Psi} \right\|_F + \|W\|_2 \left\| \widehat{\Psi} - \Psi \right\|_F \|W\|_2 \\ &\leq \left\| \widehat{W} - W \right\|_2 \left( \left\| \widehat{W} - W \right\|_2 + 2 \right) \|\Psi\|_F + \left( \left\| \widehat{W} - W \right\|_2 + \|W\|_2 \right)^2 \left\| \widehat{\Psi} - \Psi \right\|_F. \end{aligned}$$

$\square$

## F.7 Proof of Lemma E.4

By the triangle inequality and the multiplicativity of the norm with respect to the Kronecker product, (60), and (61)

$$\text{tr}(A_0) \text{tr}(B_0) \left\| \left( \widehat{W}_1^{-1} \widehat{A}^{-1} \widehat{W}_1^{-1} \right) \otimes \left( \widehat{W}_2^{-1} \widehat{B}^{-1} \widehat{W}_2^{-1} \right) \right\| \quad (66)$$

$$\begin{aligned} &\leq \|A_0^{-1}\| \|B_0^{-1}\| + \|\Delta'\| \\ &\left\| (\widehat{W}_1 \widehat{A} \widehat{W}_1) \otimes (\widehat{W}_2 \widehat{B} \widehat{W}_2) / \text{tr}(A_0) \text{tr}(B_0) \right\| \leq \|A_0\| \|B_0\| + \|\Delta\|. \end{aligned} \quad (67)$$

First following the large deviation bounds in Theorem 4.1 (see also Remark C.2), we have for  $\lambda_{A_0} \geq 3\alpha_n$  and  $\lambda_{B_0} \geq 3\beta_n$ , where  $\alpha_n \wedge \beta_n \leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{3}$ , and by (18),

$$\begin{aligned} & \left| \frac{1}{\frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2} - \frac{1}{\text{tr}(A_0)\text{tr}(B_0)} \right| = \left| \frac{\frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 - \text{tr}(A_0)\text{tr}(B_0)}{\frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 \cdot \text{tr}(A_0)\text{tr}(B_0)} \right| \\ & \leq \left| \frac{(\alpha_n \wedge \beta_n)}{\frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2} \right| \leq \frac{\alpha_n \wedge \beta_n}{\text{tr}(A_0)\text{tr}(B_0)(1 - \alpha_n \wedge \beta_n)} \\ & \leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2\text{tr}(A_0)\text{tr}(B_0)} \end{aligned} \quad (68)$$

$$\text{thus} \quad \left| \frac{\text{tr}(A_0)\text{tr}(B_0)}{\frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2} - 1 \right| \leq \frac{\alpha_n \wedge \beta_n}{1 - \alpha_n \wedge \beta_n} \leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2}. \quad (69)$$

By the triangle inequality, definition of  $\Delta$ , (67), and (69), we have

$$\begin{aligned} & \left\| \widehat{A \otimes B_0} - A_0 \otimes B_0 \right\| \\ & \leq \left| \frac{1}{\frac{1}{n} \sum_{i=1}^n \|X(i)\|_F^2} - \frac{1}{\text{tr}(A_0)\text{tr}(B_0)} \right| \left\| (\widehat{W_1} \widehat{A} \widehat{W_1}) \otimes (\widehat{W_2} \widehat{B} \widehat{W_2}) \right\| \\ & \quad + \left\| (\widehat{W_1} \widehat{A} \widehat{W_1}) \otimes (\widehat{W_2} \widehat{B} \widehat{W_2}) / \text{tr}(A_0)\text{tr}(B_0) - A_0 \otimes B_0 \right\| \\ & \leq \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2} \|A_0\| \|B_0\| + (1 + \frac{\lambda_{A_0} \wedge \lambda_{B_0}}{2}) \|\Delta\|. \end{aligned}$$

and by definition of  $\Delta'$ , (18), and (66),

$$\begin{aligned} & \left\| \widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1} \right\| \\ & = \left\| \left( \frac{1}{n} \sum_{t=1}^n \|X(t)\|_F^2 \right) \left( \widehat{W_1}^{-1} \widehat{A}^{-1} \widehat{W_1}^{-1} \right) \otimes \left( \widehat{W_2}^{-1} \widehat{B}^{-1} \widehat{W_2}^{-1} \right) - A_0^{-1} \otimes B_0^{-1} \right\| \\ & \leq \left| \frac{1}{n} \sum_{i=1}^n \|X(i)\|_F^2 - \text{tr}(A_0)\text{tr}(B_0) \right| \left\| \left( \widehat{W_1}^{-1} \widehat{A}^{-1} \widehat{W_1}^{-1} \right) \otimes \left( \widehat{W_2}^{-1} \widehat{B}^{-1} \widehat{W_2}^{-1} \right) \right\| \\ & \quad + \left\| \text{tr}(A_0)\text{tr}(B_0) \left( \widehat{W_1}^{-1} \widehat{A}^{-1} \widehat{W_1}^{-1} \right) \otimes \left( \widehat{W_2}^{-1} \widehat{B}^{-1} \widehat{W_2}^{-1} \right) - A_0^{-1} \otimes B_0^{-1} \right\| \\ & \leq (\alpha_n \wedge \beta_n) (\|A_0^{-1}\| \|B_0^{-1}\| + \|\Delta'\|) + \|\Delta'\|. \quad \square \end{aligned}$$

## G Proofs for the Flip-Flop methods

We need Lemma G.1 to define event  $\mathcal{E}_0$ , which we will use for proving Theorem 6.2 and Theorem 6.6. Lemma G.1 is a corollary of Theorem C.1, where we also denote the entry-wise error for approximating  $\rho(A_0)$  and  $\rho(B_0)$  with  $\text{diag}(A_*)^{-1/2} \tilde{A} (B_*) \text{diag}(A_*)^{-1/2}$  and  $\text{diag}(B_*)^{-1/2} \tilde{B} (A_*) \text{diag}(B_*)^{-1/2}$  by  $\lambda_{f,n}$  and  $\lambda_{m,n}$  respectively. These rates vary slightly depending on the sample size  $n$ . We use this notation throughout the rest of this section.

**Lemma G.1.** Suppose that  $n \leq \log(m \vee f)$ . Let  $d = 1 - (c/2 \wedge 1/2)$ , where  $c = \frac{\log n}{\log \log(m \vee f)}$ . On event  $\mathcal{E}_0$ ,

$$\begin{aligned} \left| \text{diag}(A_*)^{-1/2} \tilde{A}(B_*) \text{diag}(A_*)^{-1/2} - \rho(A_0) \right|_{\max} &\leq 20 \frac{\log^d \max(m, f)}{\sqrt{fn}} =: \lambda_{f,n}, \\ \left| \text{diag}(B_*)^{-1/2} \tilde{B}(A_*) \text{diag}(B_*)^{-1/2} - \rho(B_0) \right|_{\max} &\leq 20 \frac{\log^d \max(m, f)}{\sqrt{mn}} =: \lambda_{m,n} \end{aligned}$$

where  $\mathbb{P}(\mathcal{E}_0) \geq 1 - \frac{3}{\max(m, f)^2}$ .

Otherwise, let  $n > 4(C_1 \kappa^2 + C_2 \kappa) \log \max(m, f)$ , where  $\kappa > 0$ . Then, on event  $\mathcal{E}_0$ ,

$$\begin{aligned} \left| \text{diag}(A_*)^{-1/2} \tilde{A}(B_*) \text{diag}(A_*)^{-1/2} - \rho(A_0) \right|_{\max} \\ \leq 2\sqrt{C_1 + C_2/\kappa} \frac{\log^{1/2} \max(m, f)}{\sqrt{fn}} =: \lambda_{f,n} \\ \left| \text{diag}(B_*)^{-1/2} \tilde{B}(A_*) \text{diag}(B_*)^{-1/2} - \rho(B_0) \right|_{\max} \\ \leq 2\sqrt{C_1 + C_2/\kappa} \frac{\log^{1/2} \max(m, f)}{\sqrt{mn}} =: \lambda_{m,n} \end{aligned}$$

where  $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ , and  $C_2 = \sqrt{8e} \approx 7.6885$ .

## G.1 Proof of Lemma 6.1

In order to prove Lemma 6.1, we need the following auxiliary results.

**Claim G.2.** Suppose all conditions in Lemma 6.1 hold. Let  $B_1$  and  $\hat{B}$  be as defined therein. Let  $\check{B}_0 = B_1 - B_*$ , where  $B_* = \frac{\text{tr}(A_0)}{m} B_0$ . Then we have on  $\mathcal{X}_0$ ,

$$\begin{aligned} \lambda_{A_0} \left| \hat{B}^{-1} \right|_{1, \text{off}} - \frac{\alpha}{1 - \alpha} \left| \hat{B}^{-1} \right|_1 &\leq \text{tr}(\check{B}_0 B_1^{-1}) \\ &\leq \lambda_{A_0} \left| \hat{B}^{-1} \right|_{1, \text{off}} + \frac{\alpha}{1 - \alpha} \left| \hat{B}^{-1} \right|_1 =: f\tilde{\mu}. \end{aligned}$$

**Corollary G.3.** Suppose all conditions in Claim G.2 hold and let  $\lambda_{A_0} = 2\alpha/\varepsilon(1 - \alpha)$ , where  $0 < \varepsilon < 2/3$ . Then we have on  $\mathcal{X}_0$ ,

$$\left| \text{tr}(\check{B}_0 B_1^{-1}) \right| / f \leq \tilde{\mu} \leq \mu = \frac{\alpha}{1 - \alpha} \left| \rho(B_0)^{-1} \right|_1 / f + \lambda_{A_0} \left| \rho(B_0)^{-1} \right|_{1, \text{off}} / f + o(\lambda_{A_0})$$

*Proof of Lemma 6.1.* Let

$$\begin{aligned} \hat{R}_A &= \left[ \text{vec} \left\{ \tilde{S}_n^{11} \right\} \dots \text{vec} \left\{ \tilde{S}_n^{1f} \right\} \dots \text{vec} \left\{ \tilde{S}_n^{ff} \right\} \right] \\ R_A &= [b_{11} \text{vec} \{ A_0 \} \dots b_{1f} \text{vec} \{ A_0 \} \dots b_{ff} \text{vec} \{ A_0 \}] = \text{vec} \{ A_0 \} \otimes (\text{vec} \{ B_0^T \})^T. \end{aligned}$$

In order to obtain a bound on  $\left| \tilde{A}(B_1) - A_* \right|_{\max}$ , we first write  $\text{vec} \left\{ \tilde{A}(B_1) - A_* \right\}$  as follows:

$$\text{vec} \left\{ \tilde{A}(B_1) - A_* \right\} = \frac{1}{f} (\hat{R}_A - R_A) \text{vec} \{ B_*^{-1} \} + \frac{1}{f} \hat{R}_A \text{vec} \{ B_1^{-1} - B_*^{-1} \}, \quad (70)$$

where in (70) we use the following equation

$$\begin{aligned} \frac{1}{f} R_A \text{vec} \{ B_*^{-1} \} &= \frac{1}{f} \text{vec} \{ A_0 \} \otimes (\text{vec} \{ B_0^T \})^T \text{vec} \left\{ \left( \frac{\text{tr}(A_0)}{m} B_0 \right)^{-1} \right\} \\ &= \frac{m}{\text{tr}(A_0)} \text{vec} \{ A_0 \} \frac{1}{f} (\text{vec} \{ B_0^T \})^T \text{vec} \{ B_0^{-1} \} = \text{vec} \{ A_* \}. \end{aligned}$$

In order to proceed, we now compute  $\tilde{A}(B_1)$  as defined in (25), where an explicit formula for  $B_1^{-1}$  is needed. For  $\check{B}_0 := B_1 - B_*$ , we have  $\Delta^0 := B_1^{-1} - B_*^{-1} = -B_1^{-1}(B_1 - B_*)B_*^{-1} = -B_1^{-1}\check{B}_0B_*^{-1}$  and thus

$$\begin{aligned} \hat{R}_A \text{vec} \{ B_1^{-1} - B_*^{-1} \} &= \hat{R}_A (-B_1^{-1}(B_1 - B_*)B_*^{-1}) = \hat{R}_A \text{vec} \{ -B_1^{-1}\check{B}_0B_*^{-1} \} \\ &= R_A \text{vec} \{ -B_1^{-1}\check{B}_0B_*^{-1} \} + (\hat{R}_A - R_A) \text{vec} \{ \Delta^0 \}. \end{aligned}$$

We now insert the equation immediately above in (70) to obtain

$$\begin{aligned} \text{vec} \left\{ \tilde{A}(B_1) - A_* \right\} &= \frac{1}{f} (\hat{R}_A - R_A) \text{vec} \{ B_*^{-1} \} \\ &\quad + \frac{1}{f} R_A \text{vec} \{ -B_1^{-1}\check{B}_0B_*^{-1} \} + \frac{1}{f} (\hat{R}_A - R_A) \text{vec} \{ \Delta^0 \} \\ &=: U_1 + U_2 + U_3, \end{aligned} \tag{71}$$

where the matrix correspondent of each summand will be denoted by  $M_1$ ,  $M_2$ , and  $M_3$  respectively. Now for the first summand on the RHS, we have

$$\begin{aligned} U_1 &= \frac{1}{f} (\hat{R}_A - R_A) \text{vec} \{ B_*^{-1} \} = \text{vec} \left\{ \tilde{A}(B_*) - A_* \right\} \\ &= \frac{1}{f} \sum_{k=1}^f \sum_{j=1}^f \text{vec} \left\{ \tilde{S}_n^{kj} - b_{kj} A_0 \right\} B_*^{-1}(jk), \end{aligned}$$

and  $M_1 = \tilde{A}(B_*) - A_* = \frac{1}{f} \sum_{k=1}^f \sum_{\ell=1}^f \tilde{S}_n^{\ell k} (B_*^{-1})_{k\ell} - A_*$ . By Lemma G.1, we have on event  $\mathcal{E}_0$ ,

$$\forall i, j \quad |M_{1,ij}| = \left| \tilde{A}_{ij}(B_*) - a_{*,ij} \right| \leq \sqrt{a_{*,ii} a_{*,jj}} \lambda_{f,n}.$$

We now examine the second summand on the RHS of (71), where recall that  $\check{B}_0 = B_1 - B_*$ .

$$\begin{aligned} R_A \text{vec} \{ B_1^{-1} \check{B}_0 B_*^{-1} \} &= \text{vec} \{ A_0 \} \text{vec} \{ B_0^T \}^T (B_*^{-T} \otimes B_1^{-1}) \text{vec} \{ \check{B}_0 \} \\ &= \text{vec} \{ A_* \} \text{tr}(B_1^{-1} \check{B}_0). \end{aligned}$$

Then clearly on  $\mathcal{X}_0$ ,

$$\begin{aligned} U_2 &= \frac{1}{f} R_A \text{vec} \{ -B_1^{-1} \check{B}_0 B_*^{-1} \} = \frac{1}{f} \text{vec} \{ A_* \} \text{tr}(-B_1^{-1} \check{B}_0), \\ \text{and } M_2 &= A_* \frac{1}{f} \text{tr}(B_1^{-1} \check{B}_0). \end{aligned}$$

By Claim G.2 we have on event  $\mathcal{X}_0$ ,

$$\begin{aligned} |M_{2,ij}| &= |a_{*,ij}| \left| \frac{\text{tr}(\check{B}_0 B_1^{-1})}{f} \right| \\ &\leq |a_{*,ij}| \left( \lambda_{A_0} \frac{1}{f} \left| \hat{B}^{-1} \right|_{1,\text{off}} + \frac{\alpha}{1-\alpha} \frac{1}{f} \left| \hat{B}^{-1} \right|_1 \right) = |a_{*,ij}| \tilde{\mu}. \end{aligned} \tag{72}$$

Finally, we bound the third summand. Let  $\Delta^0 := B_1^{-1} - B_*^{-1}$ ,

$$U_3 := \frac{1}{f} (\widehat{R}_A - R_A) \text{vec} \{ \Delta^0 \} = \frac{1}{f} \sum_{k=1}^f \sum_{j=1}^f \text{vec} \left\{ \widetilde{S}_n^{kj} - b_{kj} A_0 \right\} \Delta_{jk}^0,$$

$$\text{and } M_3 := \frac{1}{f} \left( \sum_{k=1}^f \sum_{j=1}^f \widetilde{S}_n^{kj} \Delta_{jk}^0 - \text{tr}(B_0 \Delta^0) A_0 \right).$$

Define event  $\mathcal{E}_1$  as

$$\left| \frac{1}{f} \text{diag}(A_0)^{-1/2} \left( \sum_{k=1}^f \sum_{j=1}^f \Delta_{kj}^0 \widetilde{S}_n^{jk} \right) \text{diag}(A_0)^{-1/2} - \frac{\text{tr}(B_0 \Delta^0)}{f} \rho(A_0) \right|_{\max} \leq D' \lambda_{f,n},$$

where

$$D' = \left\| B_0^{1/2} \Delta^0 B_0^{1/2} \right\|_F / \sqrt{f} = \left\| B_0^{1/2} (B_1^{-1} - B_*^{-1}) B_0^{1/2} \right\|_F / \sqrt{f} = O \left( \frac{1}{\sqrt{f}} + \lambda_{A_0} \right),$$

and by Theorem C.1, we have  $\mathbb{P}(\mathcal{E}_1 | \mathcal{X}_0) \geq 1 - \frac{2}{(m\sqrt{f})^2}$ . To see this, we use the following bound as shown in Corollary A.3 on event  $\mathcal{X}_0$ ,

$$\begin{aligned} \left\| \widehat{B}_*^{-1} - B_*^{-1} \right\|_F &\leq \frac{2C' \lambda_{A_0}}{b_{*,\min} \varphi_{\min}^2(\rho(B_0))} \sqrt{|B_0^{-1}|_{0,\text{off}}} \vee f \quad \text{hence under (A1)} \\ D' &\leq \frac{1}{\sqrt{f}} \|B_0\|_2 \frac{2C' \lambda_{A_0}}{b_{*,\min} \varphi_{\min}^2(\rho(B_0))} \left( \sqrt{|B_0^{-1}|_{0,\text{off}}} + \sqrt{f} \right) \\ &= o \left( \frac{1}{\sqrt{f}} \right) + O(\lambda_{A_0}) = o(1). \end{aligned}$$

Thus we have under event  $\mathcal{E}_1 \cap \mathcal{X}_0$ , under (A2),

$$|M_{3,ij}| \leq \sqrt{a_{*,ii} a_{*,jj}} \frac{\text{tr}(A_0)}{m} D' \lambda_{f,n} = \sqrt{a_{*,ii} a_{*,jj}} o(\lambda_{f,n}).$$

We write the matrix correspondent of equation (71), under event  $\mathcal{X}_0 \cap \mathcal{E}_0 \cap \mathcal{E}_1$ ,

$$\begin{aligned} \left| \left( \widetilde{A}(B_1) - A_* \right)_{ij} \right| &\leq \left| \left( \widetilde{A}(B_*) - A_* \right)_{ij} \right| + \left| a_{*,ij} \frac{\text{tr}(\widetilde{B}_0 B_*^{-1})}{f} \right| + |M_{3,ij}| \\ &\leq \sqrt{a_{*,ii} a_{*,jj}} \lambda_{f,n} (1 + o(1)) + |a_{*,ij}| \widetilde{\mu}, \end{aligned}$$

where by (72),  $\widetilde{\mu} = \lambda_{A_0} \left| \widehat{B}^{-1} \right|_{1,\text{off}} / f + \frac{\alpha}{1-\alpha} \left| \widehat{B}^{-1} \right|_1 / f$  is upper bounded by  $\mu = \lambda_{A_0} \frac{1}{f} \left| \rho(B_0)^{-1} \right|_{1,\text{off}} + \frac{\alpha}{1-\alpha} \frac{1}{f} \left| \rho(B_0)^{-1} \right|_1 + o(\lambda_{A_0})$  as shown in Corollary G.3. The lemma is thus proved.  $\square$

It remains to prove Claim G.2 and Corollary G.3.

*Proof of Claim G.2.* Throughout this proof, we assume that event  $\mathcal{X}_0$  holds, and  $\lambda_{A_0} > \frac{2\alpha}{1-\alpha}$ . By definition, for  $\widehat{\Gamma}(B_0)$  be defined as in (6b),

$$\widetilde{B}(I) = \frac{1}{m} \sum_{k=1}^m \widehat{S}_n^{kk} := \frac{1}{m} \left( \sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle \right)_{i,j=1}^f = \widetilde{W}_2 \widehat{\Gamma}(B_0) \widetilde{W}_2$$

where  $\widetilde{W}_2 = \text{diag}(\widetilde{B}(I))^{1/2} \succ 0$ . We have by the KKT conditions,

$$\begin{aligned} \left| \widehat{B}_{ij} - \widehat{\Gamma}_{ij}(B_0) \right| &\leq \lambda_{A_0}, \quad \forall \widehat{B}_{ij}^{-1} = 0 \quad (\text{hence } B_{1,ij}^{-1} = 0) \\ \widehat{B}_{ij} - \widehat{\Gamma}_{ij}(B_0) &= \lambda_{A_0}, \quad \forall \widehat{B}_{ij}^{-1} > 0 \quad (\text{hence } B_{1,ij}^{-1} > 0) \\ \text{and } \widehat{B}_{ij} - \widehat{\Gamma}_{ij}(B_0) &= -\lambda_{A_0} \quad \forall \widehat{B}_{ij}^{-1} < 0 \quad (\text{hence } B_{1,ij}^{-1} < 0) \end{aligned}$$

where  $B_1 = \frac{1}{m} \widehat{W}_2 \widehat{B} \widehat{W}_2 = \widetilde{W}_2 \widehat{B} \widetilde{W}_2$ , and thus

$$\begin{aligned} \forall i, j \quad B_{1,ij} - \widetilde{B}_{ij}(I) &= \widetilde{W}_{2,ii} \left( \widehat{B}_{ij} - \widehat{\Gamma}_{ij}(B_0) \right) \widetilde{W}_{2,jj} \\ &= \begin{cases} 0 & \text{if } i = j \\ \widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj} & \text{if } B_{1,ij}^{-1} > 0 \\ -\widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj} & \text{if } B_{1,ij}^{-1} < 0 \\ \in [-\widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj}, \widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj}] & \text{if } B_{1,ij}^{-1} = 0 \end{cases}. \end{aligned}$$

Thus we have

$$\begin{aligned} \text{tr} \left( (B_1 - \widetilde{B}(I)) B_1^{-1} \right) &= \sum_{i \neq j} \widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj} \left| B_{1,ij}^{-1} \right| \\ &= \sum_{i \neq j} \widetilde{W}_{2,ii} \lambda_{A_0} \widetilde{W}_{2,jj} \left| \widetilde{W}_{2,ii}^{-1} \widehat{B}_{ij}^{-1} \widetilde{W}_{2,jj}^{-1} \right| = \lambda_{A_0} \left| \widehat{B}^{-1} \right|_{1, \text{off}}. \end{aligned}$$

The claim is proved once we show that

$$\left| \text{tr} \left( (\widetilde{B}(I) - B_*) B_1^{-1} \right) \right| \leq \left| \widehat{B}^{-1} \right|_1 \frac{\alpha}{1 - \alpha}. \quad (73)$$

Indeed, for  $\check{B}_0 = B_1 - B_* = (B_1 - \widetilde{B}(I)) + (\widetilde{B}(I) - B_*)$ , we have

$$\begin{aligned} \lambda_{A_0} \left| \widehat{B}(I)^{-1} \right|_{1, \text{off}} - \frac{\alpha}{1 - \alpha} \left| \widehat{B}(I)^{-1} \right|_1 &\leq \text{tr}(\check{B}_0 B_1^{-1}) \\ &= \text{tr} \left( (B_1 - \widetilde{B}(I)) B_1^{-1} \right) + \text{tr} \left( (\widetilde{B}(I) - B_*) B_1^{-1} \right) \\ &\leq \lambda_{A_0} \left| \widehat{B}^{-1} \right|_{1, \text{off}} + \frac{\alpha}{1 - \alpha} \left| \widehat{B}^{-1} \right|_1. \end{aligned}$$

It remains to show (73).

Before we continue, we need the following inequalities. We have on  $\mathcal{X}_0$  by Lemma 4.2 for  $\alpha := \alpha_n = \frac{m\pi_0}{\text{tr}(A_0)}$ ,

$$\begin{aligned} \forall i \neq j, \quad \left| \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{b_{ii} b_{jj} \text{tr}(A_0)}} - \rho_{ij}(B_0) \right| &\leq \alpha, \\ \text{and } \forall i = 1, \dots, f \quad \left| \frac{\frac{1}{n} \sum_{t=1}^n \|y(t)^i\|_2^2}{b_{ii} \text{tr}(A_0)} - 1 \right| &\leq \alpha, \end{aligned}$$

and hence, for all  $i$ ,

$$\begin{aligned} \frac{1}{1 + \alpha} &\leq \frac{b_{*,ii}}{\widetilde{W}_{2,ii}^2} = \frac{b_{ii} \text{tr}(A_0)}{\frac{1}{n} \sum_{t=1}^n \|y(t)^i\|_2^2} = \frac{b_{ii} \text{tr}(A_0)}{\widetilde{W}_{2,ii}^2} \leq \frac{1}{1 - \alpha}, \\ \text{and } \forall i, \quad \frac{\alpha}{1 + \alpha} &\geq 1 - \frac{b_{*,ii}}{\widetilde{W}_{2,ii}^2} = 1 - \frac{b_{ii} \text{tr}(A_0)}{\frac{1}{n} \sum_{t=1}^n \|y(t)^i\|_2^2} \geq \frac{-\alpha}{1 - \alpha}. \end{aligned}$$

Thus we have for  $\frac{1}{1+\alpha} \leq \sqrt{b_{*,ii}b_{*,jj}}/(\widetilde{W}_{2,ii}\widetilde{W}_{2,jj}) \leq \frac{1}{1-\alpha}$ ,

$$\left| \sum_{i \neq j} \widehat{B}_{ij}^{-1} \left( \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{b_{ii}b_{jj}}\text{tr}(A_0)} - \rho_{ij}(B_0) \right) \frac{\sqrt{b_{*,ii}}\sqrt{b_{*,jj}}}{\widetilde{W}_{2,ii}\widetilde{W}_{2,jj}} \right| \quad (74a)$$

$$\leq \left| \widehat{B}^{-1} \right|_{1,\text{off}} \frac{\alpha}{(1-\alpha)}, \text{ and hence}$$

$$\left| \sum_i^f \widehat{B}_{ii}^{-1} \left( 1 - \frac{b_{*,ii}}{\widetilde{W}_{2,ii}^2} \right) \right| \leq \frac{\alpha}{1-\alpha} \left| \text{diag}(\widehat{B}^{-1}) \right|_1. \quad (74b)$$

For  $\widetilde{B} := \widetilde{B}(I)$  and  $b_{*,ii}/\widetilde{W}_{2,ii}^2 = \text{tr}(A_0)b_{ii}/\widehat{W}_{2,ii}^2$ , we have

$$\begin{aligned} & \text{tr} \left( (\widetilde{B} - B_*)B_1^{-1} \right) \\ &= \text{tr} \left( \left( \widetilde{W}_2 \widehat{\Gamma}(B_0) \widetilde{W}_2 - \text{diag}(B_*)^{1/2} \rho(B_0) \text{diag}(B_*)^{1/2} \right) \widetilde{W}_2^{-1} \widehat{B}^{-1} \widetilde{W}_2^{-1} \right) \\ &= \text{tr} \left( \widehat{\Gamma}(B_0) \widehat{B}^{-1} \right) - \sum_{i,j=1}^f \rho_{ij}(B_0) \widehat{B}_{ij}^{-1} \left( \frac{\sqrt{b_{*,ii}}}{\widetilde{W}_{2,ii}} \frac{\sqrt{b_{*,jj}}}{\widetilde{W}_{2,jj}} \right) \\ &= \sum_{i \neq j} \widehat{B}_{ij}^{-1} \left( \widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0) \frac{\sqrt{b_{*,ii}}}{\widetilde{W}_{2,ii}} \frac{\sqrt{b_{*,jj}}}{\widetilde{W}_{2,jj}} \right) + \sum_i^f \widehat{B}_{ii}^{-1} \left( 1 - \frac{b_{*,ii}}{\widetilde{W}_{2,ii}^2} \right) \\ &= \sum_{i \neq j} \widehat{B}_{ij}^{-1} \left( \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{b_{ii}b_{jj}}\text{tr}(A_0)} - \rho_{ij}(B_0) \right) \frac{\sqrt{b_{*,ii}}}{\widetilde{W}_{2,ii}} \frac{\sqrt{b_{*,jj}}}{\widetilde{W}_{2,jj}} + \sum_i^f \widehat{B}_{ii}^{-1} \left( 1 - \frac{b_{*,ii}}{\widetilde{W}_{2,ii}^2} \right). \end{aligned}$$

Clearly (73) holds by taking the absolute values on both sides of the last equation, applying the triangle inequality, and then plugging in the inequalities (74a) and (74b).  $\square$

*Proof of Corollary G.3.* Throughout this proof, we assume that event  $\mathcal{X}_0$  holds. Let  $\Delta_{B_0} := \widehat{B}^{-1} - \rho(B_0)^{-1}$ . For  $\lambda_{A_0} = 2\alpha_n/\varepsilon(1-\alpha_n)$ , where  $0 < \varepsilon \leq 2/3$ , we have by (A1) and Corollary 4.4,

$$\begin{aligned} |\text{diag}(\Delta_{B_0})|_1 &\leq \sqrt{f} \|\Delta_{B_0}\|_F \leq \sqrt{f} \frac{9}{2} \frac{1+\varepsilon}{\varphi_{\min}^2(\rho(B_0))} \lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} = o(\sqrt{f}) \\ |\Delta_{B_0}|_{1,\text{off}} &\leq \sqrt{|B_0^{-1}|_{0,\text{off}}} \frac{1+\varepsilon}{1-\varepsilon} (9\lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}) / (\varphi_{\min}^2(\rho(B_0))) \\ &\leq \sqrt{|B_0^{-1}|_{0,\text{off}}} \frac{1+\varepsilon}{(1-\varepsilon)\varepsilon} \frac{\alpha_n}{1-\alpha_n} (18\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}) / (\varphi_{\min}^2(\rho(B_0))) \\ &= o\left(\sqrt{|B_0^{-1}|_{0,\text{off}}}\right) \end{aligned}$$

where  $(9\lambda_{A_0} \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}) / (\varphi_{\min}^2(\rho(B_0))) = o(1)$  by (A1) and (A2), and  $\frac{1+\varepsilon}{(1-\varepsilon)\varepsilon}$  is a constant so long as  $\varepsilon$  is bounded away from 0 and 1. Thus for  $\sqrt{|B_0^{-1}|_{0,\text{off}}} + f/f \leq 1$ , we have

$$\frac{1}{f} |\Delta_{B_0}|_1 = o(1/\sqrt{f}) + o\left(\frac{1}{f} \sqrt{|B_0^{-1}|_{0,\text{off}}}\right) = o\left(\frac{1}{f} \sqrt{|B_0^{-1}|_{0,\text{off}} + f}\right) = o(1).$$

Thus by the triangle inequality, we have for  $\lambda_{A_0} \asymp \alpha_n/(1 - \alpha_n)$ ,  $\alpha := \alpha_n$ , and  $\varepsilon < 2/3$

$$\begin{aligned} \left| \widehat{B}^{-1} \right|_{1,\text{off}} &\leq \left| \rho(B_0)^{-1} \right|_{1,\text{off}} + o\left(\sqrt{\left| B_0^{-1} \right|_{0,\text{off}}}\right), \\ \left| \widehat{B}^{-1} \right|_1 &\leq \left| \rho(B_0)^{-1} \right|_1 + |\Delta_{B_0}|_1 \leq \left| \rho(B_0)^{-1} \right|_1 + o\left(\sqrt{\left| B_0^{-1} \right|_{0,\text{off}} + f}\right), \\ \text{and hence } \quad \widetilde{\mu} &\leq \frac{\alpha}{1 - \alpha} \left| \rho(B_0)^{-1} \right|_1 / f + \lambda_{A_0} \left| \rho(B_0)^{-1} \right|_{1,\text{off}} / f + o(\lambda_{A_0}). \end{aligned}$$

The corollary thus holds.  $\square$

## G.2 Proof of Theorem 6.2

First we observe that (34) follows from (33) immediately given that  $\widetilde{\eta} = \lambda_{f,n}(1 + o(1)) + \widetilde{\mu} < \eta$ . We now show that (33) follows from (30). Indeed, we have for all  $i, j$ , on event  $\mathcal{A}_1$ ,

$$\begin{aligned} \left| \widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \right| &:= \left| \frac{\widetilde{A}_{ij}(B_1)}{\widetilde{A}_{ii}^{1/2}(B_1)\widetilde{A}_{jj}^{1/2}(B_1)} - \rho_{ij}(A_0) \right| \\ &= \left| \frac{\widetilde{A}_{ij}^{1/2}(B_1)/\sqrt{a_{*,ii}a_{*,jj}} - \rho_{ij}(A_0)}{(\widetilde{A}_{ii}^{1/2}(B_1)/\sqrt{a_{*,ii}})(\widetilde{A}_{jj}^{1/2}(B_1)/\sqrt{a_{*,jj}})} \right| \\ &\quad + \left| \frac{\rho_{ij}(A_0)}{(\widetilde{A}_{ii}^{1/2}(B_1)/\sqrt{a_{*,ii}})(\widetilde{A}_{jj}^{1/2}(B_1)/\sqrt{a_{*,jj}})} - \rho_{ij}(A_0) \right| \\ &\leq \frac{\lambda_{f,n}(1 + o(1)) + \widetilde{\mu} |\rho_{ij}(A_0)|}{1 - \widetilde{\eta}} + |\rho_{ij}(A_0)| \left| \frac{1}{1 - \widetilde{\eta}} - 1 \right| \\ &= \frac{\lambda_{f,n}(1 + o(1))}{1 - \widetilde{\eta}} + |\rho_{ij}(A_0)| \frac{\widetilde{\eta} + \widetilde{\mu}}{1 - \widetilde{\eta}} \\ &= \frac{\lambda_{f,n}(1 + o(1))}{1 - \widetilde{\eta}} + |\rho_{ij}(A_0)| \frac{\lambda_{f,n}(1 + o(1)) + 2\widetilde{\mu}}{1 - \widetilde{\eta}} \\ &= \frac{2}{1 - \widetilde{\eta}} (\lambda_{f,n}(1 + o(1)) + |\rho_{ij}(A_0)| \widetilde{\mu}) \leq \frac{2\widetilde{\eta}}{1 - \widetilde{\eta}} \end{aligned}$$

where we used the fact that on  $\mathcal{A}_1$ , by (30),

$$\begin{aligned} \forall i, \quad \left| \frac{\widetilde{A}(B_1)_{ii}}{a_{*,ii}} - 1 \right| &\leq \widetilde{\eta} \text{ and hence } \frac{\widetilde{A}(B_1)_{ii}^{1/2}}{\sqrt{a_{*,ii}}} \geq \sqrt{1 - \widetilde{\eta}} \\ \text{and } \forall i \neq j, \quad \left| \frac{\widetilde{A}(B_1)_{ij}}{\sqrt{a_{*,ii}a_{*,jj}}} - \frac{a_{*,ij}}{\sqrt{a_{*,ii}a_{*,jj}}} \right| &\leq \lambda_{f,n}(1 + o(1)) + \frac{|a_{*,ij}|}{\sqrt{a_{*,ii}a_{*,jj}}} \widetilde{\mu}. \quad \square \end{aligned}$$

## G.3 Proof of Corollary 6.4

Throughout this proof, we assume that event  $\mathcal{A}_1$  holds and  $m \leq f$ . Clearly event  $\mathcal{T}(A_0)$  holds for sample correlation matrix  $\widehat{\Gamma}(A_0)$  for  $\delta_{n,f} = \frac{2\widetilde{\eta}}{1 - \widetilde{\eta}} \leq \frac{2\eta}{1 - \eta} \asymp \lambda_{f,n}$  on event  $\mathcal{A}_1$ , where  $\delta_{f,n} \sqrt{\left| A_0^{-1} \right|_{0,\text{off}} \vee 1} = o(1)$



by (A1). We have by Theorem 4.3, on event  $\mathcal{A}_1$ ,

$$\left\| \hat{A}(B_1)^{-1} - \rho(A_0)^{-1} \right\|_F \leq 9(1 + \varepsilon) \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1 / (2\varphi_{\min}^2(\rho(A_0)))}, \quad (75)$$

$$\text{and } \left\| \hat{A}(B_1) - \rho(A_0) \right\|_F \leq 9(1 + \varepsilon) \kappa(\rho(A_0))^2 \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}. \quad (76)$$

Let  $\tilde{W}_1 = \text{diag}(\tilde{A}(B_1))^{1/2}$  and  $W = \text{diag}(A_*)^{1/2} = \text{diag}(\sqrt{a_{*,11}}, \dots, \sqrt{a_{*,mm}}) = \sqrt{\frac{m}{\text{tr}(A_0)}} \text{diag}(A_0)^{1/2}$ .

We have for all  $i$ , by (30),  $|\tilde{W}_{1,ii}^2 - a_{*,ii}| \leq a_{*,ii} \tilde{\eta}$ . Then the following holds on  $\mathcal{A}_1$ ,

$$\left\| \tilde{W}_1 - W \right\|_2 \leq (\sqrt{1 + \tilde{\eta}} - 1) \vee (1 - \sqrt{1 - \tilde{\eta}}) a_{*,\max}^{1/2} \leq \sqrt{a_{*,\max}} \tilde{\eta}, \quad (77)$$

$$\begin{aligned} \left\| \tilde{W}_1^{-1} - W^{-1} \right\|_2 &\leq \left( \frac{\sqrt{1 + \tilde{\eta}} - 1}{\sqrt{1 + \tilde{\eta}}} \vee \frac{1 - \sqrt{1 - \tilde{\eta}}}{\sqrt{1 - \tilde{\eta}}} \right) \frac{1}{\sqrt{a_{*,\min}}} \\ &\leq \frac{\tilde{\eta}}{\sqrt{1 - \tilde{\eta}}} \frac{1}{\sqrt{a_{*,\min}}}. \end{aligned} \quad (78)$$

By Proposition F.1, (76), and (77), and for  $\tilde{\eta} < \lambda_{B_1}(1 - \tilde{\eta})/2$ ,  $\eta < 1/4$ , and  $18 > C > 9$ ,

$$\begin{aligned} \left\| \hat{A}_* - A_* \right\|_2 &= \left\| \tilde{W}_1 \hat{A}(B_1) \tilde{W}_1 - \text{diag}(A_*)^{1/2} \rho(A_*) \text{diag}(A_*)^{1/2} \right\|_2 \\ &\leq (\tilde{\eta} + 2) \tilde{\eta} a_{*,\max} \|\rho(A_0)\|_2 + C \lambda_{B_1} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} a_{*,\max} (1 + \tilde{\eta})^2 \\ &\leq \lambda_{B_1} \frac{(1 - \tilde{\eta})(\tilde{\eta} + 2)}{2} a_{*,\max} \|\rho(A_0)\|_2 + C \lambda_{B_1} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} a_{*,\max} (1 + \tilde{\eta})^2 \\ &\leq 2C \lambda_{B_1} a_{*,\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} \quad \text{and} \\ \left\| \hat{A}_* - A_* \right\|_F &\leq \lambda_{B_1} \sqrt{m} a_{*,\max} \|\rho(A_0)\|_2 + C \lambda_{B_1} a_{*,\max} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} (1 + \tilde{\eta})^2 \\ &\leq 2C a_{*,\max} \lambda_{B_1} \kappa(\rho(A_0))^2 \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m}. \end{aligned}$$

Similarly, by Proposition F.1, (75), and (78), we have for  $\tilde{\eta} \leq \eta < 1/4$ , and  $9 > C > 9/2$ ,

$$\begin{aligned} \left\| \Delta^1 \right\|_2 &:= \left\| \hat{A}_*^{-1} - A_*^{-1} \right\|_2 = \left\| \tilde{W}_1^{-1} \hat{A}(B_1)^{-1} \tilde{W}_1^{-1} - \text{diag}(A_*)^{-1/2} \Psi^{-1} \text{diag}(A_*)^{-1/2} \right\|_2 \\ &\leq \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left( \tilde{\eta} + 2\sqrt{1 - \tilde{\eta}} \right) / (\varphi_{\min}(\rho(A_0)) a_{*,\min}) + \frac{C \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{\varphi_{\min}^2(\rho(A_0)) a_{*,\min}} \left( 1 + \frac{\tilde{\eta}}{\sqrt{1 - \tilde{\eta}}} \right)^2 \\ &\leq 2C \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} / (\varphi_{\min}^2(\rho(A_0)) a_{*,\min}) \quad \text{and} \\ \left\| \Delta^1 \right\|_F &:= \left\| \hat{A}_*^{-1} - A_*^{-1} \right\|_F \leq 2C \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}} \vee m} / (\varphi_{\min}^2(\rho(A_0)) a_{*,\min}). \quad \square \end{aligned}$$

#### G.4 Proof of Theorem 6.6 and Lemma 6.5

*Proof of Theorem 6.6.* Given Lemma 6.5, the proof follows exactly the same arguments as Theorem 6.2, and hence is omitted.  $\square$

*Proof of Lemma 6.5.* Let

$$\begin{aligned}\widehat{R}_B &= \left[ \text{vec} \left\{ \widehat{S}_n^{11} \right\} \dots \text{vec} \left\{ \widehat{S}_n^{1m} \right\} \dots \text{vec} \left\{ \widehat{S}_n^{mm} \right\} \right], \\ R_B &= [a_{11} \text{vec} \{ B_0 \} \dots a_{1m} \text{vec} \{ B_0 \} \dots a_{mm} \text{vec} \{ B_0 \}] = \text{vec} \{ B_0 \} \otimes (\text{vec} \{ A_0^T \})^T.\end{aligned}$$

Let  $A_1 = \widehat{A}_*$ . For convenience, we write  $\Delta^1 = A_1^{-1} - A_*^{-1} = \widehat{A}_*^{-1} - A_*^{-1}$  and  $\check{A}_1 := A_1 - A_*$ . We have

$$\Delta^1 = A_1^{-1} - A_*^{-1} = -A_1^{-1}(A_1 - A_*)A_*^{-1} = -A_1^{-1}\check{A}_1A_*^{-1}. \quad (79)$$

First we write

$$\begin{aligned}& \frac{1}{m} \text{vec} \left\{ \widetilde{B}(A_1) \right\} - \frac{1}{m} R_B \text{vec} \left\{ A_*^{-1} \right\} = \frac{1}{m} \widehat{R}_B \text{vec} \left\{ A_1^{-1} \right\} - \text{vec} \{ B_* \} \\ &= \frac{1}{m} (\widehat{R}_B - R_B) \text{vec} \left\{ A_*^{-1} \right\} + \frac{1}{m} R_B \text{vec} \left\{ \Delta^1 \right\} + \frac{1}{m} (\widehat{R}_B - R_B) \text{vec} \left\{ \Delta^1 \right\} \\ &:= V_1 + V_2 + V_3,\end{aligned} \quad (80)$$

where the matrix correspondent of each summand will be denoted by  $M_1$ ,  $M_2$ , and  $M_3$  respectively. Now for the first summand on the RHS, we have

$$\begin{aligned}V_1 &= \frac{1}{m} (\widehat{R}_B - R_B) \text{vec} \left\{ A_*^{-1} \right\} = \text{vec} \left\{ \widetilde{B}(A_*) - B_* \right\} \\ &= \frac{1}{f} \sum_{k=1}^f \sum_{j=1}^f \text{vec} \left\{ \widehat{S}_n^{kj} - a_{kj} B_0 \right\} A_*^{-1}(jk)\end{aligned}$$

and  $M_1 = \widetilde{B}(A_*) - B_* = \frac{1}{m} \sum_{k=1}^f \sum_{\ell=1}^f \widehat{S}_n^{\ell k} (A_*^{-1})_{k\ell} - B_*$ .

By Lemma G.1, we have on event  $\mathcal{E}_0$ ,

$$|M_{1,ij}| = \left| \widetilde{B}_{ij}(A_*) - B_{*,ij} \right| \leq \sqrt{b_{*,ii} b_{*,jj}} \lambda_{m,n}.$$

We now examine the second summand on the RHS of (80), where recall that  $\check{A}_1 = \widehat{A}_* - A_*$ . Now by (79),

$$\begin{aligned}V_2 &= \frac{1}{m} R_B \text{vec} \left\{ -A_1^{-1} \check{A}_1 A_*^{-1} \right\} = \frac{1}{m} \text{vec} \{ B_0 \} \text{vec} \left\{ A_0^T \right\}^T (A_*^{-T} \otimes A_1^{-1}) \text{vec} \left\{ -\check{A}_1 \right\} \\ &= \frac{1}{m} \text{vec} \{ B_0 \} \text{tr}(-A_0 A_1^{-1} \check{A}_1 A_*^{-1}) = \frac{1}{m} \text{vec} \{ B_* \} \text{tr}(-A_1^{-1} \check{A}_1).\end{aligned}$$

Hence  $M_2 = B_* \text{tr}(-A_1^{-1} \check{A}_1)/m$ . From Claim G.4, we have on event  $\mathcal{A}_1$ ,

$$|M_{2,ij}| = |b_{*,ij}| |\text{tr}(-\check{A}_1 A_1^{-1})| / m \leq |b_{*,ij}| \left( \lambda_{B_1} \left| \widehat{A}^{-1} \right|_{1,\text{off}} + \frac{\widetilde{\eta}}{1 - \widetilde{\eta}} \left| \widehat{A}^{-1} \right|_1 \right) = |b_{*,ij}| \widetilde{\xi}.$$

Finally, we bound the third summand. Let  $\Delta^1$  be as in (79). By definition,

$$V_3 = \frac{1}{m} (\widehat{R}_B - R_B) \text{vec} \left\{ \Delta^1 \right\} = \frac{1}{m} \sum_{k=1}^m \sum_{j=1}^m \text{vec} \left\{ \widehat{S}_n^{kj} - a_{kj} B_0 \right\} \Delta_{jk}^1$$

and  $M_3 = \frac{1}{m} \sum_{k=1}^m \sum_{j=1}^m \widehat{S}_n^{kj} \Delta_{jk}^1 - \frac{\text{tr}(A_0 \Delta^1)}{m} B_0$ . Define event  $\mathcal{E}_2$  as

$$\begin{aligned}& \left| \frac{1}{m} \text{diag}(B_0)^{-1/2} \left( \sum_{k=1}^m \sum_{j=1}^m \Delta_{kj}^1 \widehat{S}_n^{jk} \right) \text{diag}(B_0)^{-1/2} - \frac{\text{tr}(A_0 \Delta^1)}{m} \rho(B_0) \right|_{\max} \leq D_4 \lambda_{m,n}, \\ & \text{where } D_4 = \left\| A_0^{1/2} \Delta^1 A_0^{1/2} \right\|_F / \sqrt{m} \leq \|A_0\|_2 \|A_1^{-1} - A_*^{-1}\|_F / \sqrt{m} = o(1).\end{aligned}$$

Then under event  $A_1$ , we have by proof of Theorem C.1,  $\mathbb{P}(\mathcal{E}_2|\mathcal{A}_1) \geq 1 - \frac{2}{(m\sqrt{f})^2}$ . To see this, we have by Corollary 6.4, on event  $\mathcal{A}_1$ ,

$$\begin{aligned} D_4 &\leq \frac{1}{\sqrt{m}} \|A_0\|_2 \|A_1^{-1} - A_*^{-1}\|_F \\ &\leq \frac{1}{\sqrt{m}} \|A_0\|_2 2C\lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}}} \vee m/(a_{*,\min}\varphi_{\min}^2(\rho(A_0))) \\ &\asymp \frac{1}{\sqrt{m}} \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}}} + \lambda_{B_1} = o\left(\frac{1}{\sqrt{m}}\right) + O(\lambda_{B_1}) = o(1) \end{aligned}$$

where note that under (A1), (A2) and (A3), we have for  $0 < \varepsilon_1 < 1$ ,

$$\lambda_{B_1} = \frac{2\tilde{\eta}}{\varepsilon_1(1-\tilde{\eta})} \asymp \lambda_{f,n} + \lambda_{m,n} = O(\lambda_{B_0}) \rightarrow 0 \quad \text{and} \quad \lambda_{B_1} \sqrt{|A_0^{-1}|_{0,\text{off}}} = o(1).$$

We have under event  $\mathcal{E}_2 \cap \mathcal{A}_1$ ,

$$|M_{3,ij}| \leq \sqrt{b_{ii}b_{jj}} D_4 \lambda_{m,n} = \sqrt{b_{*,ii}b_{*,jj}} \frac{m}{\text{tr}(A_0)} D_4 \lambda_{m,n} = \sqrt{b_{*,ii}b_{*,jj}} o(\lambda_{m,n}).$$

The lemma is thus proved by plugging in the bounds on  $|M_{1,ij}|$ ,  $|M_{2,ij}|$ , and  $|M_{3,ij}|$ . On  $\mathcal{A}_1 \cap \mathcal{E}_2$ , we have

$$\begin{aligned} \left| \left( \tilde{B}(A_1) - B_* \right)_{ij} \right| &\leq \left| \left( \tilde{B}(A_*) - B_* \right)_{ij} \right| + \left| b_{*,ij} \frac{\text{tr}(\tilde{A}_1 A_*^{-1})}{m} \right| + |M_{3,ij}| \\ &\leq \sqrt{b_{*,ii}b_{*,jj}} \lambda_{m,n} (1 + o(1)) + |b_{*,ij}| \tilde{\xi} \end{aligned}$$

by (80), where  $\tilde{\xi} = \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}}/m + \frac{\tilde{\eta}}{1-\tilde{\eta}} \left| \hat{A}^{-1} \right|_1/m$  is bounded by  $\xi$  as shown in Corollary G.5.  $\square$

In order to prove Lemma 6.5, we need the following auxiliary results.

**Claim G.4.** Suppose all conditions in Lemma 6.5 hold. Let  $\tilde{W}_1 = \text{diag}(\tilde{A}(B_1))^{1/2}$ . Let  $A_1 = \hat{A}_* := \tilde{W}_1 \hat{A}(B_1) \tilde{W}_1$ , where  $\hat{A}(B_1)$  is as obtained in Step 2 with  $\lambda_{B_1}$  chosen to be lower bounded by  $2\tilde{\eta}/(1-2\tilde{\eta})$  where  $\tilde{\eta} := \lambda_{f,n}(1+o(1)) + \tilde{\mu}$ . Let  $\tilde{A}_1 := A_1 - A_*$ . Then we have on  $\mathcal{A}_1$  for  $\hat{A} := \hat{A}(B_1)$ ,

$$\begin{aligned} \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}} - \frac{\tilde{\eta}}{1-\tilde{\eta}} \left| \hat{A}^{-1} \right|_1 &\leq \\ \text{tr}(\tilde{A}_1 \hat{A}_1^{-1}) &\leq \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}} + \frac{\tilde{\eta}}{1-\tilde{\eta}} \left| \hat{A}^{-1} \right|_1. \end{aligned} \tag{81}$$

**Corollary G.5.** Suppose all conditions in Claim G.4 hold. Let  $0 < \varepsilon_1 < 1$ , and

$$\lambda_{B_1} = \frac{2\tilde{\eta}}{\varepsilon_1(1-\tilde{\eta})} = O(\lambda_{B_0}) \text{ where } \tilde{\eta} = \lambda_{f,n}(1+o(1)) + \tilde{\mu}$$

is as defined as in (34). Clearly  $\frac{\tilde{\eta}}{1-\tilde{\eta}} = \frac{\varepsilon_1}{2} \lambda_{B_1} \leq \lambda_{B_1}/3$ . Then we have on  $\mathcal{A}_1$ , for  $\hat{A} := \hat{A}(B_1)$ ,

$$\frac{|\text{tr}(\tilde{A}_1 \hat{A}_1^{-1})|}{m} \leq \tilde{\xi} \leq \lambda_{B_1} \frac{|\rho(A_0)^{-1}|_{1,\text{off}}}{m} + \frac{\tilde{\eta}}{1-\tilde{\eta}} \frac{|\rho(A_0)^{-1}|_1}{m} + o(\lambda_{B_1}) \leq \xi$$

where  $\xi$  is as defined in (37).

*Proof of Claim G.4.* Throughout this proof, we assume that event  $\mathcal{A}_1$  holds. On  $\mathcal{A}_1$ , we have by (30),

$$\forall i \neq j, \quad \left| \frac{\tilde{A}_{ij}(B_1)}{\sqrt{a_{*,ii}a_{*,jj}}} - \frac{a_{*,ij}}{\sqrt{a_{*,ii}a_{*,jj}}} \right| \leq \lambda_{f,n}(1 + o(1)) + \frac{|a_{*,ij}|}{\sqrt{a_{*,ii}a_{*,jj}}} \tilde{\mu}, \quad (82a)$$

$$\text{and } \forall i, \quad 1 - \tilde{\eta} \leq \frac{\tilde{W}_{1,ii}^2}{a_{*,ii}} \leq 1 + \tilde{\eta},$$

$$\text{and hence } \frac{1}{\sqrt{1 + \tilde{\eta}}} \leq \frac{\sqrt{a_{*,ii}}}{\tilde{W}_{1,ii}} \leq \frac{1}{\sqrt{1 - \tilde{\eta}}}, \quad (82b)$$

as  $\text{diag}(\tilde{A}(B_1)) = \tilde{W}_1^2$ , and  $\left| \tilde{W}_{1,ii}^2/a_{*,ii} - 1 \right| \leq \tilde{\eta} = \lambda_{f,n}(1 + o(1)) + \tilde{\mu} < \eta$ . By the KKT conditions we obtain for  $A_1 = \hat{A}_* = \tilde{W}_1 \hat{A}(B_1) \tilde{W}_1$  and  $\tilde{A}(B_1) = \tilde{W}_1 \hat{\Gamma}(A_0) \tilde{W}_1$ , where  $\tilde{W}_1 \succ 0$ ,

$$\begin{aligned} \left| \hat{A}_{ij}(B_1) - \hat{\Gamma}_{ij}(A_0) \right| &\leq \lambda_{B_1}, \forall \hat{A}_{ij}^{-1}(B_1) = 0 \quad (\text{hence } A_{1,ij}^{-1} = 0) \\ \hat{A}_{ij}(B_1) - \hat{\Gamma}_{ij}(A_0) &= \lambda_{B_1}, \quad \forall \hat{A}_{ij}^{-1}(B_1) > 0 \quad (\text{hence } A_{1,ij}^{-1} > 0) \\ \text{and } \hat{A}_{ij}(B_1) - \hat{\Gamma}_{ij}(A_0) &= -\lambda_{B_1} \quad \forall \hat{A}_{ij}^{-1}(B_1) < 0 \quad (\text{hence } A_{1,ij}^{-1} < 0), \end{aligned}$$

and hence for all  $i, j$ ,

$$\begin{aligned} A_{1,ij} - \tilde{A}_{ij}(B_1) &= \tilde{W}_{1,ii} \left( \hat{A}_{ij}(B_1) - \hat{\Gamma}_{ij}(A_0) \right) \tilde{W}_{1,jj} \\ &= \begin{cases} 0 & \text{if } i = j \\ \tilde{W}_{1,ii} \lambda_{B_1} \tilde{W}_{1,jj} & \text{if } A_{1,ij}^{-1} > 0 \\ -\tilde{W}_{1,ii} \lambda_{B_1} \tilde{W}_{1,jj} & \text{if } A_{1,ij}^{-1} < 0 \\ \in [-\tilde{W}_{1,ii} \lambda_{B_1} \tilde{W}_{1,jj}, \tilde{W}_{1,ii} \lambda_{B_1} \tilde{W}_{1,jj}] & \text{if } A_{1,ij}^{-1} = 0 \end{cases}. \end{aligned}$$

Thus we have

$$\begin{aligned} \text{tr} \left( (A_1 - \tilde{A}(B_1)) A_1^{-1} \right) &= \sum_{i \neq j} \tilde{W}_{1,ii} \lambda_{B_1} \tilde{W}_{1,jj} \left| \tilde{W}_{1,ii}^{-1} \hat{A}_{ij}(B_1)^{-1} \tilde{W}_{1,jj}^{-1} \right| \\ &= \lambda_{B_1} \left| \hat{A}(B_1)^{-1} \right|_{1,\text{off}}. \end{aligned}$$

The claim is proved if we show that for  $\hat{A} := \hat{A}(B_1)$ , where  $\hat{A}(B_1)$  is obtained in Step 2,

$$\left| \text{tr} \left( (\tilde{A}(B_1) - A_*) A_1^{-1} \right) \right| \leq \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \hat{A}^{-1} \right|_1 \quad (83)$$

so that for  $\check{A}_1 = A_1 - A_*$ , we have

$$\begin{aligned} \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}} / m - \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \hat{A}^{-1} \right|_1 / m &\leq \text{tr}(\check{A}_1 A_1^{-1}) / m \\ &= \frac{1}{m} \text{tr} \left( (A_1 - \tilde{A}(B_1)) A_1^{-1} \right) + \frac{1}{m} \text{tr} \left( (\tilde{A}(B_1) - A_*) A_1^{-1} \right) \\ &\leq \lambda_{B_1} \left| \hat{A}^{-1} \right|_{1,\text{off}} / m + \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \hat{A}^{-1} \right|_1 / m. \end{aligned}$$

It remains to show (83). Denote by  $\tilde{A} := \tilde{A}(B_1) = \tilde{W}_1 \hat{\Gamma}(A_0) \tilde{W}_1$ . Let  $\text{diag}(A_*)^{1/2} = \text{diag}(\sqrt{a_{*,11}}, \dots, \sqrt{a_{*,mm}}) = \sqrt{\frac{m}{\text{tr}(A_0)}} \text{diag}(A_0)^{1/2}$ . We have

$$\begin{aligned}
& \text{tr} \left( (\tilde{A} - A_*) A_1^{-1} \right) \\
&= \text{tr} \left( \left( \tilde{W}_1 \hat{\Gamma}(A_0) \tilde{W}_1 - \text{diag}(A_*)^{1/2} \rho(A_0) \text{diag}(A_*)^{1/2} \right) \tilde{W}_1^{-1} \hat{A}^{-1} \tilde{W}_1^{-1} \right) \\
&= \text{tr} \left( \hat{\Gamma}(A_0) \hat{A}^{-1} \right) - \sum_{i,j=1}^m \rho_{ij}(A_0) \hat{A}_{ij}^{-1} \left( \frac{\sqrt{a_{*,ii}}}{\tilde{W}_{1,ii}} \frac{\sqrt{a_{*,jj}}}{\tilde{W}_{1,jj}} \right) \\
&= \sum_{i \neq j} \hat{A}_{ij}^{-1} \left( \hat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0) \frac{\sqrt{a_{*,ii}}}{\tilde{W}_{1,ii}} \frac{\sqrt{a_{*,jj}}}{\tilde{W}_{1,jj}} \right) + \sum_i \hat{A}_{ii}^{-1} \left( 1 - \frac{a_{*,ii}}{\tilde{W}_{1,ii}^2} \right) \\
&= \sum_{i \neq j} \hat{A}_{ij}^{-1} \left( \frac{\tilde{A}_{ij}}{\sqrt{a_{*,ii} a_{*,jj}}} - \rho_{ij}(A_0) \right) \frac{\sqrt{a_{*,ii}}}{\tilde{W}_{1,ii}} \frac{\sqrt{a_{*,jj}}}{\tilde{W}_{1,jj}} + \sum_i \hat{A}_{ii}^{-1} \left( 1 - \frac{a_{*,ii}}{\tilde{W}_{1,ii}^2} \right), \tag{84}
\end{aligned}$$

where by (82a) and (82b), we have

$$\begin{aligned}
& \left| \sum_i \hat{A}_{ii}^{-1} \left( 1 - \frac{a_{*,ii}}{\tilde{W}_{1,ii}^2} \right) \right| \leq \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \text{diag}(\hat{A}^{-1}) \right|_1, \\
& \text{and } \left| \sum_{i \neq j} \hat{A}_{ij}^{-1} \left( \frac{\tilde{A}_{ij}(B_1)}{\sqrt{a_{*,ii} a_{*,jj}}} - \rho_{ij}(A_0) \right) \frac{\sqrt{a_{*,ii}}}{\tilde{W}_{1,ii}} \frac{\sqrt{a_{*,jj}}}{\tilde{W}_{1,jj}} \right| \\
& \leq \sum_{i \neq j} \frac{|\hat{A}_{ij}^{-1}|}{1 - \tilde{\eta}} (\lambda_{f,n}(1 + o(1)) + \tilde{\mu} |\rho_{ij}(A_0)|) \leq \frac{\tilde{\eta}}{1 - \tilde{\eta}} \left| \hat{A}^{-1} \right|_{1, \text{off}} / m.
\end{aligned}$$

Clearly (83) holds by inserting the above inequalities in (84).  $\square$

*Proof of Corollary G.5.* Throughout this proof, we assume that event  $\mathcal{A}_1$  holds. Let  $\Delta_{A_1} = \hat{A}(B_1)^{-1} - \rho(A_0)^{-1}$ . We have for  $\lambda_{B_1} = 2\tilde{\eta}/\varepsilon_1(1 - \tilde{\eta})$ , where  $0 < \varepsilon_1 < 1$ ,

$$\begin{aligned}
|\text{diag}(\Delta_{A_1})|_1 &\leq \sqrt{m} \|\Delta_{A_1}\|_F \\
&\leq \sqrt{m} 9(1 + \varepsilon_1) \lambda_{B_1} \frac{\sqrt{|A_0^{-1}|_{0, \text{off}} \vee 1}}{2\varphi_{\min}^2(\rho(A_0))} = o(\sqrt{m})
\end{aligned}$$

where  $\lambda_{B_1} \sqrt{|A_0^{-1}|_{0, \text{off}} \vee 1} = o(1)$  by (A1), given that  $\lambda_{B_1} \asymp \frac{\eta}{1 - \eta} = O(\lambda_{B_0})$ . Now for the non-diagonal part of  $\Delta_{A_1}$ , for  $0 < \varepsilon_1 < 1$ , we have by Corollary 4.4,

$$\begin{aligned}
|\Delta_{A_1}|_{1, \text{off}} &\leq \sqrt{|A_0^{-1}|_{0, \text{off}}} \frac{1 + \varepsilon_1}{1 - \varepsilon_1} 9\lambda_{B_1} \sqrt{|A_0^{-1}|_{0, \text{off}} \vee 1 / (\varphi_{\min}^2(\rho(A_0)))} \\
&\leq \sqrt{|A_0^{-1}|_{0, \text{off}}} \frac{1 + \varepsilon_1}{(1 - \varepsilon_1)\varepsilon_1} \frac{18\tilde{\eta}}{1 - \tilde{\eta}} \sqrt{|A_0^{-1}|_{0, \text{off}} \vee 1 / (\varphi_{\min}^2(\rho(A_0)))} \\
&= o\left(\sqrt{|A_0^{-1}|_{0, \text{off}}}\right)
\end{aligned}$$

where  $\frac{1+\varepsilon_1}{(1-\varepsilon_1)\varepsilon_1}$  is bounded so long as  $\varepsilon_1$  is bounded away from 0 and 1. Hence

$$\begin{aligned}
\frac{1}{m} |\Delta_{A_1}|_1 &= o(1/\sqrt{m}) + o\left(\frac{1}{m} \sqrt{|A_0^{-1}|_{0,\text{off}}}\right) = o\left(\frac{1}{m} \sqrt{|A_0^{-1}|_{0,\text{off}} + m}\right) = o(1) \\
|\widehat{A}(B_1)^{-1}|_{1,\text{off}}/m &\leq |\rho(A_0)^{-1}|_{1,\text{off}}/m + o\left(\frac{1}{m} \sqrt{|A_0^{-1}|_{0,\text{off}}}\right) \\
&= |\rho(A_0)^{-1}|_{1,\text{off}}/m + o(1), \text{ and hence} \\
|\widehat{A}(B_1)^{-1}|_1/m &\leq |\rho(A_0)^{-1}|_1/m + |\Delta_{A_1}|_1/m = |\rho(A_0)^{-1}|_1/m + o(1)
\end{aligned}$$

where clearly  $\frac{\sqrt{|A_0^{-1}|_{0,\text{off}} + m}}{m} \leq 1$ . We now insert the inequalities above in (81):

$$\begin{aligned}
\tilde{\xi} &= \lambda_{B_1} \frac{1}{m} |\widehat{A}^{-1}|_{1,\text{off}} + \frac{\tilde{\eta}}{1-\tilde{\eta}} \frac{1}{m} |\widehat{A}^{-1}|_1 \\
&\leq \lambda_{B_1} \left( \frac{1}{m} |\rho(A_0)^{-1}|_{1,\text{off}} + o\left(\frac{1}{m} \sqrt{|A_0^{-1}|_{0,\text{off}}}\right) \right) \\
&\quad + \frac{\tilde{\eta}}{1-\tilde{\eta}} \left( |\rho(A_0)^{-1}|_1/m + o\left(\frac{1}{m} \sqrt{|A_0^{-1}|_{0,\text{off}} + m}\right) \right) \\
&\leq \frac{\tilde{\eta}}{1-\tilde{\eta}} |\rho(A_0)^{-1}|_1/m + \lambda_{B_1} |\rho(A_0)^{-1}|_{1,\text{off}}/m + o(\lambda_{B_1}).
\end{aligned}$$

The other bounds follow from the fact that  $\tilde{\eta} \leq \eta$ . The corollary thus holds.  $\square$